UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA EN TELECOMUNICACIÓN

SONIDO E IMAGEN

PROYECTO FIN DE CARRERA

EXTRACCIÓN DE PARÁMETROS ACÚSTICOS PARA LA DISCRIMINACIÓN ENTRE MÚSICA Y VOZ

AUTOR: RAÚL PÉREZ LÓPEZ

TUTORA: ASCENSIÓN GALLARDO ANTOLÍN

Diciembre 2009

Agradecimientos

A Ascensión, mi tutora y guía en este proyecto de fin de carrera, por su comprensión y paciencia infinita, sin lo cual no habríamos llegado a concluir este trabajo en ningún momento.

Estaba Jesús en cierta ocasión junto al lago de Genesaret y la gente se agolpaba para oír la palabra de Dios. Vio entonces dos barcas en la orilla del lago; los pescadores habían desembarcado y estaban lavando las redes. Subió a una de las barcas, que era de Simón, y le pidió que la separase un poco de tierra. Se sentó y estuvo enseñando a la gente desde la barca. Cuando terminó de hablar, dijo a Simón:

- Rema lago adentro y echad vuestras redes para pescar.

Simón respondió:

- Maestro, hemos estado toda la noche faenando sin pescar nada, pero puesto que tú lo dices, echaré las redes.

Lo hicieron y capturaron una gran cantidad de peces. Como las redes se rompían, hicieron señas a sus compañeros de la otra barca para que vinieran a ayudarlos. Vinieron y llenaron las dos barcas, hasta el punto de que casi se hundían. Al verlo, Simón Pedro cayó a los pies de Jesús diciendo:

- Apártate de mí, Señor, que soy un pecador.

Pues tanto él como sus hombres estaban sobrecogidos de estupor ante la cantidad de peces que habían capturado; e igualmente Santiago y Juan, hijos de Zebedeo, que eran compañeros de Simón. Entonces Jesús dijo a Simón:

- No temas, desde ahora serás pescador de hombres.

Y después de llevar las barcas a tierra, dejaron todo y lo siguieron.

Lc 5,1-11

Resumen

Este Proyecto de Fin de Carrera se enmarca dentro de un conjunto de líneas de investigación sobre el procesado de registros de audio que lleva a cabo el Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid.

El trabajo se ha centrado en una de esas líneas, teniendo como objetivo evaluar el rendimiento de un sistema de clasificación de audio con diferentes parámetros utilizados habitualmente para la caracterización de la señal de audio, como son la energía, la tasa de cruces por cero y la frecuencia fundamental.

Se han realizado una serie de experimentos con el objeto de determinar las características acústicas que permiten una mejor clasificación del audio en las dos clases propuestas para este proyecto: habla y música instrumental.

Índice de contenidos

Capítulo 1. Introducción

1.1 Resumen

El Proyecto de Fin de Carrera que se presenta a continuación engloba la extracción de características de audio básicas para la posterior clasificación del audio en dos categorías: voz y música instrumental.

Para llevar a cabo esta clasificación se ha trabajado a partir de características de audio de la energía, la tasa de cruces por cero y la frecuencia fundamental, con la intención de caracterizar en la medida de lo posible las muestras de audio para su posterior clasificación.

Para ellos se llevan a cabo una serie de pasos: extracción de las características de audio por tramas o por segmentos, interpretación de los histogramas obtenidos para la elección de los parámetros a utilizar, entrenamiento del clasificador y la última etapa que es la clasificación por medio del clasificador lineal o cuadrático. Una vez realizado todo el proceso todo el proceso se analizan los resultados obtenidos para obtener conclusiones al respecto.

1.2 Objetivo

La intención de este proyecto es hallar un conjunto de características propias de la señal de audio que permita diferenciar de la mejor forma posible los diferentes tipos de archivos de audio. Para este proyecto se quieren diferenciar en dos clases: habla y música instrumental.

Pero, ¿cómo hacer esa diferenciación?, ¿qué características de estos archivos de audio nos permiten diferenciar un tipo de información respecto de otro? En este proyecto trataremos de descubrir qué características acústicas permiten una discriminación más eficaz entre estos tipos de audio.

Por tanto, el objetivo de este proyecto es determinar experimentalmente qué parámetros de la señal de audio nos permitirán una mejor discriminación entre habla y música instrumental.

Para la clasificación del habla se ha estado utilizando de forma exitosa los parámetros Cepstrales, MFCC. En los últimos años, la llegada del MPEG7 ha abierto una serie de nuevas posibilidades para la etiquetación y clasificación de audio gracias a los descriptores que desde ese estándar se proponen, parámetros de envolvente espectral, ASE y ASP. En este proyecto se propone la utilización de otras características básicas del audio (energía, tasa de cruces por cero y frecuencia fundamental) extraídas tanto a nivel de trama como a nivel de segmento.

1.3 Estructura del proyecto

La memoria de este Proyecto de Fin de Carrera se estructura en 8 capítulos y un anexo. A continuación se presenta el contenido de cada uno de ellos como referencia global del conjunto total del proyecto.

Capítulo 1. Introducción

Este capítulo ofrece una visión global de todo lo que es el proyecto: resumen del mismo, objetivos y una estructura general. De tal modo que permite al lector situarse frente a lo que va a leer.

Capítulo 2. Estado del Arte

Se presenta el contexto de la situación actual frente a la clasificación de archivos de audio y se ofrece una visión global de las técnicas de discriminación entre voz y música instrumental.

Capítulo 3. Caracterización paramétrica de música y voz

Se describe el conjunto de parámetros que se han desarrollado para la realización de este proyecto: energía, tasa de cruces por cero y frecuencia fundamental. El desarrollo del proyecto no solo se ha centrado en cada uno de estos tres parámetros, también se han utilizado los logaritmos de los mismos y variaciones concretas de ellos, que se explican a lo largo de este capítulo.

Capítulo 4. Sistema de clasificación de registros de audio

En este capítulo se ofrece una visión general del clasificador lineal y clasificador cuadrático, los dos clasificadores utilizados para la categorización en dos clases de las pistas de audio.

Capítulo 5. Resultados experimentales

Se engloba todos los resultados obtenidos de forma experimental al aplicar sobre los clasificadores cada una de las características acústicas descritas en el capítulo 3 o diversas combinaciones de ellas. Se muestran los resultados y se comparan entre ellos para obtener conclusiones.

Capítulo 6. Conclusiones y líneas futuras

Presenta la síntesis de los resultados obtenidos tras la realización del proyecto, destacándose los objetivos alcanzados así como las problemáticas no resueltas. Finalmente se indican las líneas por las que podrían ir futuras investigaciones para la mejora de la aplicación implementada.

Capítulo 7. Presupuesto

Este capítulo presenta una estimación del presupuesto final necesario para el desarrollo del proyecto. Para el cálculo del mismo se tienen en cuenta los costes del material empleado y los costes de los honorarios de los desarrolladores.

Capítulo8. Bibliografía/Referencias

Se muestran las referencias bibliográficas de las que se ha hecho uso a lo largo del proyecto.

Anexo. Resultados experimentales

Se presentan de forma ordenada todos y cada uno de los resultados obtenidos en la realización del proyecto. Se muestran resultados que en el capítulo 5 no aparecen por no sobrecargarlo en exceso.

Capítulo 2.Estado del arte

2.1 Introducción

En los últimos años, se han propuesto diferentes sistemas para la discriminación automática de voz y música. Saunders propuso un clasificador en tiempo real de voz y música para ser utilizados en los receptores de radio, para el control automático del contenido de Canales de radio FM. El reconocimiento automático del habla (ASR) es importante en la emisión de noticias para desactivar el reconocedor de voz durante la parte no hablada de la secuencia de audio. Recientemente, Scheirer y Slaney y Williams y Ellis desarrollaron y evaluaron diferentes sistemas de clasificación de música y voz para ASR de pistas de audio.

La tarea de segmentación de una secuencia de audio y clasificación de cada segmento como música o voz, es importante en una serie de aplicaciones de procesamiento, incluido la extracción de información de programas de radio, el reconocimiento automático del habla y la codificación de voz a tasas binarias bajas. Por ejemplo, en un sistema de navegación digital de radio, un oyente interesado sólo en la música sería capaz de saltar de forma automática segmentos del habla. Incluso sería posible obtener el perfil de la estación de radio, es decir, decidir si la radio esta orientada a voz o a música.

Otra aplicación que puede obtenerse de diferenciar voz y música es la codificación de audio a tasas binarias bajas. Tradicionalmente, se han diseñado codificadores diferentes para voz y para música. Generalmente, los codificadores de voz funcionaban mejor con el habla, y los codificadores de audio mejor con la música. En las nuevas aplicaciones multimedia, como las de Internet, la señal puede contener segmentos de voz y música, por lo que se han invertido muchos esfuerzos en el diseño de un codificador universal que codifique de forma aceptable tanto la palabra como la música. Sin embargo, esto no es un problema trivial. Un enfoque alternativo es el diseño de un

codificador multi-modo que pueda adaptarse a diferentes señales. El módulo adecuado en cada momento es seleccionado usando la salida del clasificador de música y voz. Este enfoque ya ha sido empleado en el codificador de parámetros de MPEG-4 y más recientemente en el codificador de audio propuesto por Ramprashad, y en el codificador de la banda ancha de música y voz. [2.1]

Una aplicación multimedia emergente es la extracción de contenido a partir de audio y vídeo, donde la clasificación de audio es una parte importante de tales sistemas. La clasificación automática eliminaría la subjetividad inherente en el proceso de clasificación y aceleraría el proceso de recuperación de información. Zhang y Kuo desarrollaron un sistema de recuperación basado en contenido de audio que clasifica las señales como habla, música o ruido. Minami propuso un enfoque basado en audio para la indexación de vídeo. Un clasificador de música y voz se utiliza para ayudar a los usuarios a navegar por una base de datos de vídeo. [2.2]

El problema de distinguir las señales de voz de otras señales de audio se ha convertido en un paso previo al reconocimiento automático del habla (ASR), ya que los sistemas se aplican a más situaciones reales del mundo multimedia, como la transcripción automática de noticias, en el que la voz está normalmente intercalada con segmentos de música y de otros ruidos de fondo. Reconocedores de voz estándar intentan realizar el reconocimiento en todos los segmentos de entrada y, naturalmente, se producen grandes errores con esa señal de entrada que mezcla música y voz. Por lo tanto, es necesaria una pre-fase que diferencie segmentos en voz y no voz, lo que ofrecerá una mejor precisión en la etapa de reconocimiento. Esto también tiene el beneficio de reducir la carga computacional en general, ya que el sistema de reconocimiento de voz sólo se activa para los segmentos necesarios.

Más en general, la segmentación de audio podría permitir el uso de modelos acústicos ASR entrenados en particular para ciertas condiciones acústicas, tales como banda ancha (entrada de micrófono de alta calidad) frente al ancho de banda estrecha del teléfono convencional, voz de hombre

frente a la de mujer, etc., lo que mejora de forma global el rendimiento del sistema resultante. Por último, esta segmentación también puede ser diseñada para proporcionar interesante información adicional, como la diferenciación entre hablantes y la identidad de los mismos (permitiendo, por ejemplo, una indexación automática y la recuperación de todas las intervenciones de una misma persona), al igual que información sintáctica (final de la frase, marcas de puntuación, etc.).

La estructura básica de un sistema de clasificación de audio se caracteriza en tres etapas claramente diferenciadas: la extracción de características con la que el clasificador va a trabajar; una segunda etapa es el entrenamiento del clasificador; y por último la etapa de test, en la que se comprueba los resultados que ofrece dicho clasificador. A continuación se describe brevemente las principales técnicas utilizadas en cada etapa.

2.2 Parametrización de señales de audio

Uno de los problemas en el diseño de un clasificador de la señal es la selección de un conjunto de características adecuadas, aquellas que mejor capturan la estructura temporal y espectral de las señales. Muchas de estas funciones para la discriminación entre música y voz ya se han sugerido, incluyendo información de cruce por cero, de la energía, el tono, los coeficientes Cepstrales, línea espectral de frecuencias, la energía de la modulación a 4 Hz, amplitud, y las características perceptivas como el timbre y el ritmo (Sheirer y Slaney, 1997, Saunders, 1996; Carey, et al., 1999; El-Maleh et al., 2000; Parris et al., 1999). [2.3]

Los sistemas de clasificación existentes de música y voz suelen usar características que se prolongan a lo largo del tiempo, tales como las variaciones. Del mismo modo que la tonalidad y el tono también se han combinado en varios diseños. Normalmente, estas características se estiman para segmentos de audio de 0.5-5 segundos, puesto que un oyente humano puede discriminar con facilidad entre la voz y una señal de música al escuchar un segmento corto (es decir, unos pocos segundos) de una señal de audio.

2.3 Clasificación y segmentación de señales de audio

Otro problema en el diseño del sistema es la selección de un algoritmo de clasificación. Se han utilizado para este propósito diferentes clasificadores como el criterio de información bayesiano (BIC) (Chen y Gopalkrishnan, 1998), la relación de probabilidad de Gauss (GLR) (Sheirer y Slaney, 1997; Saunders, 1996; Parris et al., 1999, Williams y Ellis, 1999), clasificador cuadrático (QGC), (El-Maleh et al., 2000), clasificador según el vecino más próximo (clasificador Sheirer y Slaney, 1997; El-Maleh et al. 2000) y modelos ocultos de Markov (HMM) (Zhang y Kuo, 1999). [2.3]

La esencia del esquema de categorización se basó en si las tareas de segmentación y clasificación son tratadas de forma conjunta o por separado. Esta última hipótesis ha sido adoptada por la mayoría de las propuestas, es decir, un algoritmo de segmentación divide el flujo de audio en segmentos y en una segunda etapa, un sistema de clasificación binaria se aplica en cada segmento. Como resultado, la investigación se ha centrado en el potencial de los diversos sistemas de clasificación para etiquetar cada segmento.

Hoy en día, un algoritmo basado en el BIC (Criterio de Información Bayesiano, Chen y Gopalkrishnan, 1998) es quizás el más utilizado en la técnica de segmentación de audio. Se supone que el vector de secuencias de características acústicas sufre un proceso gaussiano, y mide la probabilidad de que dos ventanas consecutivas fueran generadas por dos procesos en lugar de por un único proceso. La técnica BIC es útil para la detección de cambios de audio generales, ya que no requiere ningún conocimiento previo sobre las clases acústicas particulares. En las aplicaciones reales, plantea una serie de problemas prácticos, tales como la alta carga computacional y la necesidad de ajustar un parámetro de umbral (k) para optimizar el rendimiento.

En ocasiones se utiliza la entropía y características dinámicas estimadas en la salida del perceptrón multicapa (MLP, denominada MLP primaria) usado en híbridos regulares HMM/MLP de sistema de reconocimiento de voz. Dependiendo de los datos presentados en la entrada de la MLP primaria, estas

características presentan diferentes propiedades y se puede utilizar en un segundo estado (estado del habla / no voz) de sistemas HMM, donde la densidad de probabilidad de un estado se calcula por cualquiera de los modelos de mezcla de Gaussianas (GMM) o secundaria MLP. Este enfoque tiene dos ventajas:

1. utilizar las características que han demostrado tener propiedades discriminantes entre voz y música, y
2. ser un umbral libre, estrategia de toma de decisiones a nivel mundial.

En el mismo marco, también se investiga la utilización de una medida de confianza para mejorar el rendimiento de la discriminación del sistema. Esta medida puede utilizarse para mejorar la exactitud de la discriminación mediante la eliminación de segmentos de baja confianza. Además, esa medida de confianza podría ser utilizada en el marco de mezcla de música y voz, donde es deseable determinar la cantidad de la palabra o de música presente en la señal de audio, en lugar de simplemente proporcionar los límites de la segmentación dura.

Capítulo 3. Caracterización paramétrica de música y voz

3.1 Introducción

En este capítulo se va a describir el conjunto de los parámetros que han sido desarrollados para la realización de este proyecto de un clasificador de audio en dos categorías (música y voz).

Uno de los aspectos más importantes a la hora de desarrollar todo el conjunto del proyecto ha sido seleccionar los parámetros o características acústicas que de una forma más clara nos pudieran diferenciar los dos tipos de clases de audio con las que se trabaja (archivos de música instrumental y archivos de voz). Así pues y en base a experiencias previas en el campo de los reconocedores de habla y clasificadores de audio se estimó que algunos de los parámetros más destacables para lograr clasificar el audio serían la energía, la tasa de cruces por cero (ZCR) y la frecuencia fundamental (F_0).

El desarrollo del proyecto no solo se ha centrado en cada uno de estos tres parámetros, sino que también se han utilizado los logaritmos de los mismos y variaciones concretas de estos parámetros, que se explicarán a lo largo de este capítulo.

Con este conjunto de variaciones de los parámetros se consigue una mayor variedad de información, un conjunto más amplio con el que trabajar, lo que previsiblemente permitirá seleccionar el sub-parámetro que en cada caso sea más favorable para lograr así una óptima clasificación.

3.2 Procedimiento

La visión general del procedimiento que se ha llevado a cabo para la extracción de los parámetros es la siguiente:

- Extracción del parámetro en sí por medio de su fórmula en concreto.
- Cálculo, en su caso, de las variaciones del parámetro (logaritmo, cálculo por tramas, cálculo por segmentos…)
- Obtención del histograma del parámetro y sus variaciones.
- Cálculo de la distancia KL Simétrica (Kullback-Leibler) entre mismos parámetros de ambas clases de audio.

3.3.1 Extracción del parámetro y sus variaciones

La extracción de cada uno de los parámetros se lleva a cabo por medio de la programación de su fórmula (ver los apartados de cada una de las características dentro de este capítulo).

Para ello se ha utilizado el software matemático MATLAB, el cual ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M), que permitirá de forma fácil obtener un resultado realmente válido para el clasificador final que se quiere lograr.

Del mismo modo la extracción de características a partir de las tres básicas (Energía, ZCR y F0) se ha desarrollado con este lenguaje de programación.

3.3.2 Obtención del histograma

Una vez que se obtienen una determinada característica acústica se realiza el histograma de la misma, que servirá para tener, de forma visual, una representación de los valores obtenidos.

Esta representación gráfica facilitará la comparación de cada una de las características entre los dos tipos de clases con las que se trabaja. Es decir, gracias al histograma, se podrá ver de forma gráfica la tendencia que tienen los valores obtenidos y así poder realizar una rápida valoración de la similitud o no similitud de una misma característica de cada una de las clases consideradas.

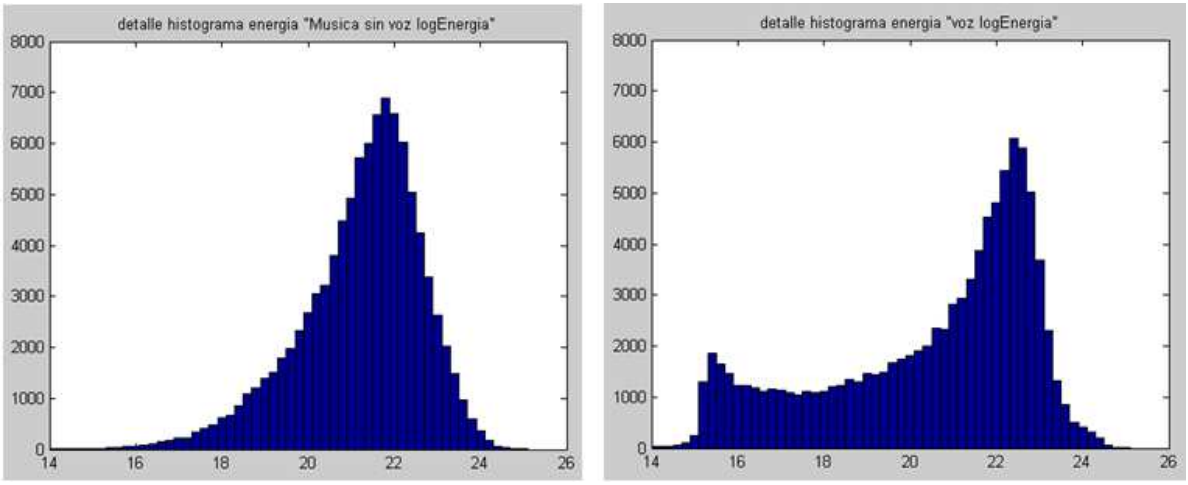

Figura 1 Histogramas de la logEnergia para ambos tipos de clases.

Como se observa en la *Figura 1* gracias a los histogramas se puede ver de forma gráfica la tendencia de cada uno de los parámetros con los que se trabaja (es este caso en concreto la log-energía) y así valorar de forma intuitiva su relevancia para el clasificador de audio.

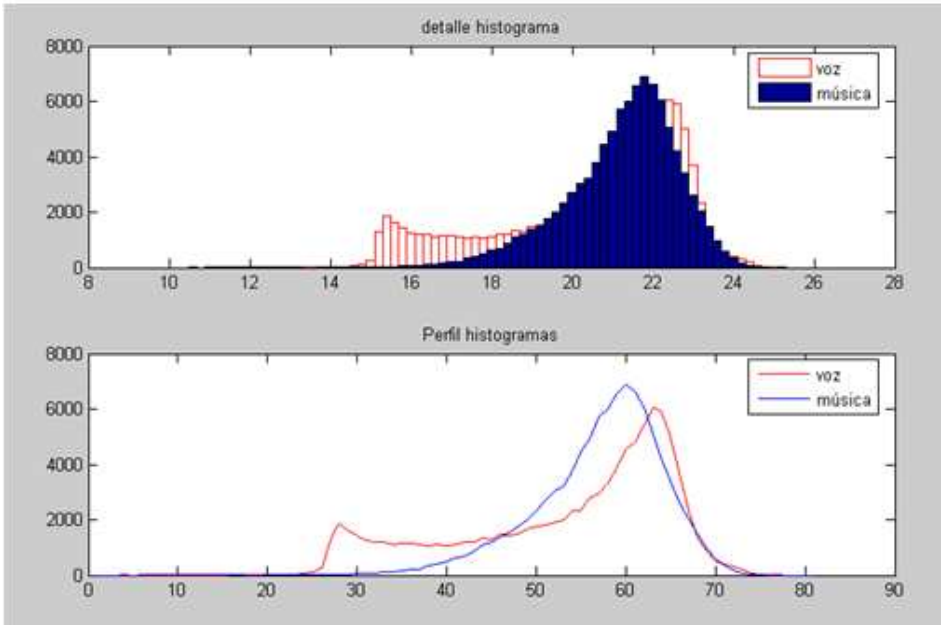

Figura 2: Histogramas *Figura 1* para una mejor comparativa

Como se observa en la *Figura 2* realizar una comparativa superpuesta facilita el entendimiento de forma fácil de la paridad o disparidad de una misma característica entre las dos categorías.

3.2.3 Distancia KL Simétrica (Kullback-Leibler)

Una vez calculado el histograma, el siguiente paso es el cálculo de la distancia KLs (Distancia Kullback-Leibler Simétrica), que indica de forma numérica lo que dos características se parecen entre sí.

A continuación procedemos a la definición de la distancia de Kullback-Leibler [3.1]. Sea *x* una variable aleatoria que toma valores continuos en el intervalo Ω y sean *f(x)* y *g(x)* las densidades de probabilidad de dos procesos aleatorios. Se define la entropía relativa de *g* respecto a *f*, el número de Kullback-Leibler o la discriminación entre los dos procesos como:

$$H(f,g) = \int_{x\in\Omega} f(x) log \frac{f(x)}{g(x)}\, dx$$

Ecuación 1: Entropía relativa

La entropía relativa entre dos distribuciones de probabilidad es no negativa, siendo nula únicamente si las dos distribuciones son idénticas. De este modo, se puede considerar como una medida de la divergencia entre dos distribuciones de probabilidad.

Cuando se consideran distribuciones de tipo Gaussiano, la entropía relativa se puede calcular como una función de las medias y las desviaciones típicas. Sean $p_1(x)$ y $p_2(x)$ dos distribuciones de probabilidad Gaussianas de medias μ_1 y μ_2 y desviaciones típicas σ_1 y σ_2, respectivamente. La entropía relativa de $p_2(x)$ respecto a $p_1(x)$ se calcula como:

$$H\big(p_1(x), p_2(x)\big) = \int_{-\infty}^{+\infty} p_1(x) log \frac{p_1(x)}{p_2(x)}\, dx$$

Ecuación 2: Entropía relativa de ($p_2(x)$, $p_1(x)$)

Considerando $H\big(p_1(x), p_2(x)\big)$ como el valor esperado de la función $log(p_1(x)/p_2(x))$ sobre $p_1(x)$, es decir, $E_1[log(p_1(x)/p_2(x))]$, se obtiene:

$$H\big(p_1(x), p_2(x)\big) = E_1\left[log\frac{\sigma_2}{\sigma_1} + \frac{(x-\mu_2)^2}{2\sigma_2^2} + \frac{(x-\mu_1)^2}{2\sigma_1^2}\right]$$

Ecuación 3: Entropía relativa (desarrollo)

Finalmente, desarrollando cada uno de los términos de la ecuación anterior se llega a que la entropía relativa se puede calcular como:

$$H\big(p_1(x), p_2(x)\big) = \frac{1}{2}\left[log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} - 1\right]$$

Ecuación 4: Entropía relativa

Dada la definición de entropía relativa de la *ecuación 1*, resulta claro que no es una medida de distancia puesto que no es simétrica (de ahí que también adquiera el nombre de desviación o dispersión). Sin embargo, se puede redefinir una distancia simetrizada como:

$$KLs\left(p_1(x), p_2(x)\right) = H\big(p_1(x), p_2(x)\big) + H\big(p_2(x), p_1(x)\big)$$

Ecuación 5: Entropía relativa simetrizada

Que para distribuciones de tipo Gaussiano como las que se consideran en este proyecto, adquiere la forma:

$$KLs\left(p_1(x), p_2(x)\right) = \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + (\mu_1-\mu_2)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) - 2\right]$$

Ecuación 6: Entropía relativa simetrizada Ecuaciones Gaussinas

De este modo se obtiene un valor numérico que indica la disparidad entre las dos distribuciones que se consideren en cada caso; siendo este valor cero cuando las dos distribuciones sean la misma y pudiendo tomar cualquier otro valor positivo para distribuciones diferentes.

3.2.4 Bolsa de datos utilizada

El conjunto de datos que se utiliza para realizar todos los procesos a lo largo de este proyecto consta de 800 ficheros de audio solo voz y 800 ficheros de música instrumental.

Cada uno de los ficheros es una división de un fichero inicial de una duración superior. Los ficheros actuales tienen una duración temporal de 1.5 segundos, tiempo que permitirá extraer de forma adecuada las características que se requieren para el desarrollo del proceso.

3.3 Energía / logEnergia

La energía no es un estado físico real, es un número escalar que se le asigna a un estado de un sistema físico, es decir, la energía es una herramienta matemática de una propiedad de los sistemas físicos. Por tanto, la energía es un valor arbitrario de cada una de las muestras que se van a analizar. [3.2]

En el caso de las señales de audio, se suele asemejar energía de la señal a la intensidad sonora de la misma, siendo señales con mucha energía aquellas que tienen un sonido muy alto, ya que la propagación del sonido involucra propagación de energía sin propagar materia, en forma de ondas mecánicas que se propagan a través de la materia sólida, líquida o gaseosa. Por tanto y para el caso que aquí se analiza tener está idea en la cabeza será más o menos válida a la hora de analizar y comparar señales de audio.

3.3.1 Cálculo de la energía

El cálculo de la energía se realiza a partir del valor al cuadrado de la señal acotada en un tipo de ventana en concreto.

Las ventanas son funciones matemáticas usadas en el análisis de señales para evitar discontinuidades al principio y al final de cada uno de los bloques analizados. Así pues una ventana se utiliza cuando es necesario acotar de forma voluntaria el tamaño de la señal con la que se va a trabajar, ya que el cálculo de cualquier valor (en este caso la energía) solo es posible si se trabaja con un número finito de valores. [3.3]

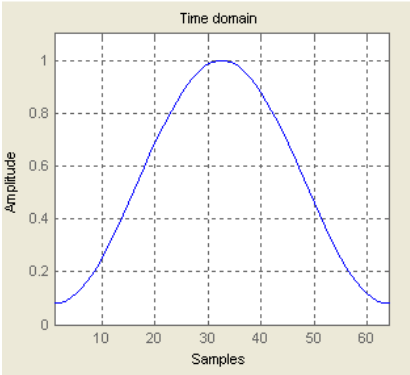

Figura 3: Ventana Hamming

En este caso, la ventana con la que se procede al acotamiento de la señal es una ventana tipo Hamming (*figura 3*), que suaviza los valores de los extremos evitando así discontinuidades bruscas entre segmentos consecutivos. Su fórmula es la siguiente, en la que N indica la longitud de la ventana en muestras [3.4].

$$w(n) = \begin{cases} 0.54 - 0.46\, cos\left({}^{2\pi n}/_{(N-1)}\right), & 0 \leq n \leq N-1 \\ 0, & en\ caso\ contrario \end{cases}$$

Ecuación 7: Fórmula de la ventana Hamming

Podría parecer que si lo único que se quiere conseguir es acotar la señal de audio la ventana más apropiada para ello sea una ventana rectangular, que seleccionará un número determinado de muestras con las que se calculará la energía. El problema es que una ventana rectangular modifica demasiado la señal en el dominio de la frecuencia, por tanto se utilizan otro tipo de ventanas, que aún modificando en mayor medida la señal en el dominio temporal la mantienen suficientemente invariante en el dominio frecuencial.

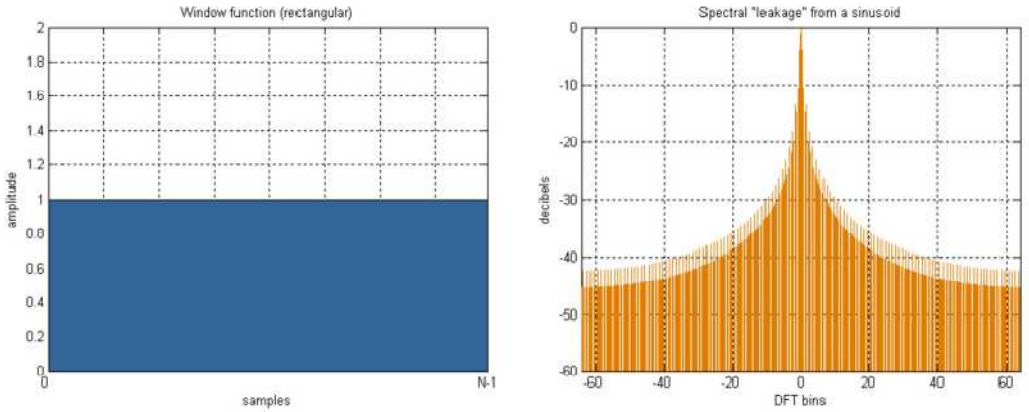

Figura 4: Ventana Rectangular en el dominio del tiempo y de la frecuencia

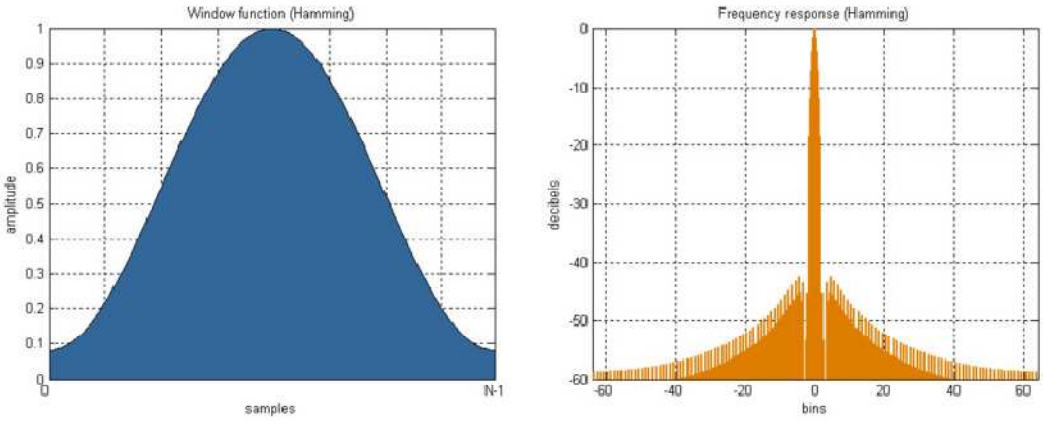

Figura 5: Ventana Hamming en el dominio del tiempo y de la frecuencia

Como se aprecia en la *figura 4* y en la *figura 5*, la ventana rectangular no modifica la señal en el dominio temporal, mientras que la Hamming sí que lo hace. Por el contrario, la ventana Hamming en el dominio frecuencial se acerca más a una *delta*, que es el caso ideal en el que el dominio frecuencial no es modificado, o sea convolucionar una señal cualquiera en el dominio frecuencial por una *delta* mantiene a la señal invariante.

Del mismo modo, para evitar discontinuidades, la ventana no se desplaza cada vez todo el tamaño de su longitud, si no que se desplaza la mitad de su longitud para así irse superponiendo y lograr una ponderación más adecuada del valor que se está calculando.

Siendo *s* la señal de audio y *w* la ventana Hamming con la que se va a hacer el acotado de la señal, se define la energía como:

$$E = \sum_{i=1}^{N} s(i)^2 \cdot w(i)$$

Ecuación 8: Fórmula de la energía

3.3.2 Idea intuitiva sobre la energía

Antes de analizar de forma cuantitativa cada una de las características con las que vamos a trabajar es conveniente tener una idea de partida del resultado aproximado que se espera obtener.

Para comenzar hay que tener claro las dos categorías con las que se va a trabajar para poder tener una idea inicial válida del resultado que se puede lograr. Así pues en el desarrollo de este proyecto se trabaja con muestras de audio que contienen solo voz hablada y muestras de audio de música exclusivamente instrumental.

Como punto de partida inicial se puede considerar que las muestras que van a tener una energía mayor serán las de música instrumental, ya que la música se mantiene más constante a lo largo del tiempo, con grandes variaciones de dinámica pero sin apenas silencios, por lo que tiene un comportamiento muy regular. Por el contrario, la voz es mucho más irregular, ya que entre las palabras pueden aparecer silencios (zonas con energía prácticamente nula). Además, la voz sufre grandes variaciones de energía entre los propios fonemas, lo que crea que en un estudio en conjunto la energía de la voz sea inferior a la de la música.

Si se analizan las representaciones gráficas de unas señales aleatorias de música y de voz se puede intentar discernir lo anteriormente explicado:

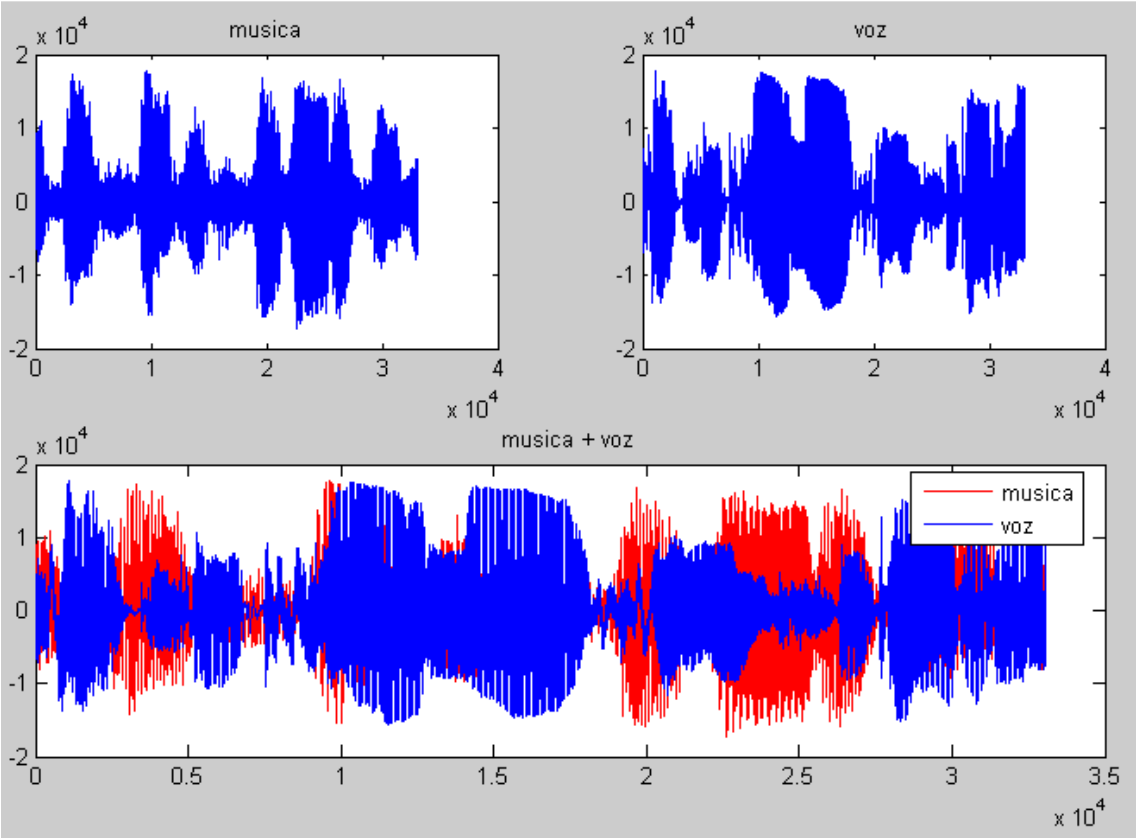

Figura 6: Representación gráfica de una señal de música, una de voz y ambas señales superpuestas

Quizá gráficamente lo único que se puede observar a primera vista en la *figura 6* es la mayor regularidad de la señal de música frente a la señal de voz. Pero si se dedica un poco más de tiempo al análisis visual de estas gráficas podemos observar que en la señal de música no hay ningún silencio, o apenas los hay, mientras que el la señal de voz tenemos zonas con gran cantidad de energía y otras zonas con energía prácticamente nula.

Con un análisis más detallado de las gráficas se puede observar lo siguiente:

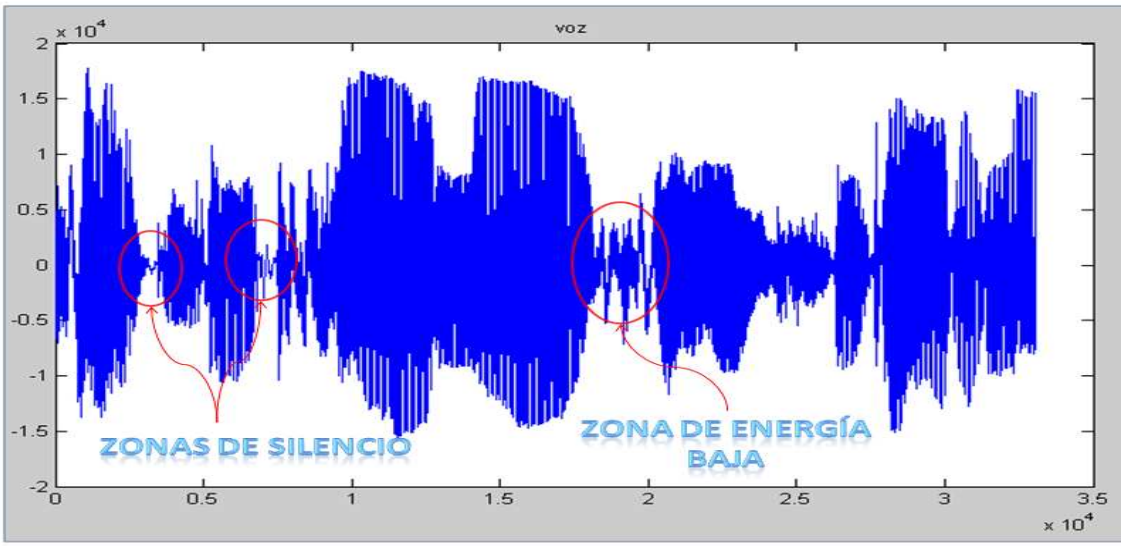

Figura 7: Detalles de las características de la señal de voz

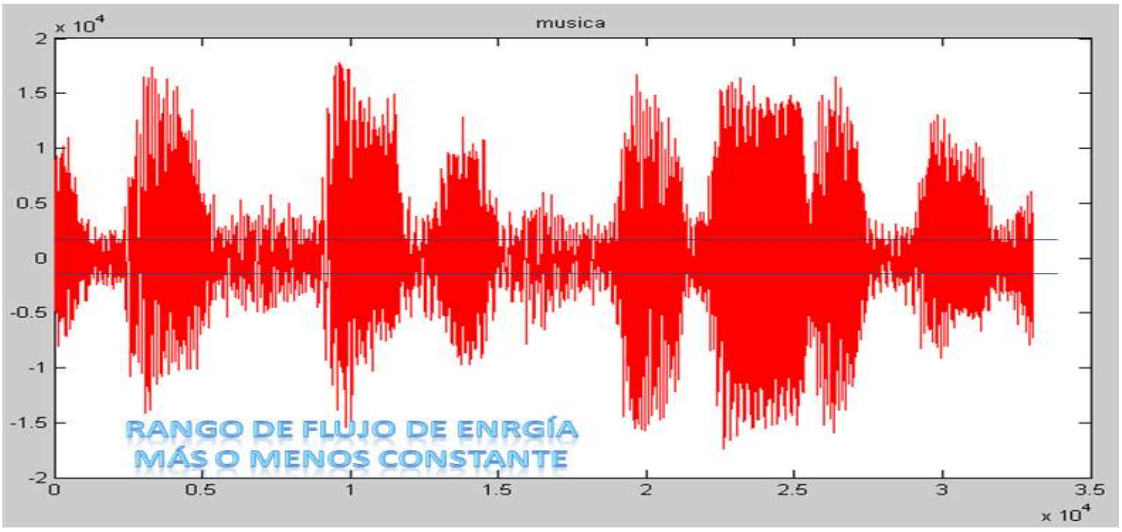

Figura 8: Detalles de las características de la señal de música

Como se observa en los detalles remarcados en la *Figura 7* la señal de voz presenta zonas con poca cantidad de energía o energía prácticamente nula, mientras que la señal de música (*Figura 8*) se mantiene con un rango mucho más constante a lo largo del tiempo.

Por todo lo anteriormente expuesto, se considera de partida que las señales de música tendrán una energía superior a las señales de voz. Ahora bien, en el

apartado siguiente, dentro de este mismo capítulo, se realiza un análisis matemático cuantitativo de las señales y se mostrará la veracidad o no de lo anteriormente expuesto.

3.3.3 Histogramas de la energía

En este apartado se procede al análisis de cada uno de los histogramas obtenidos para el parámetro *Energía*. Los cálculos sobre la energía que se han realizado son los siguientes: *energía y logEnergia por trama*, *media y varianza de la energía y logEnergía en segmentos de un segundo*, y *cálculo del porcentaje de tramas con energía baja.*

El cálculo por trama se realiza sobre el conjunto total de los valores de cada una de los ficheros de audio. Mientras que el cálculo de características segmentales se realiza sobre segmentos de audio de 1 segundo de duración. Es decir, en el caso, por ejemplo, de la media de la energía, se calcula el vector de valores de la energía en un tramo de un segundo y una vez que se tiene ese vector se hace su media, obteniendo así un único valor para cada segundo procesado. Así pues, por cada segundo que se analiza se tiene un valor medio de la energía.

La ponderación en segmentos de un segundo se realiza ya que el oído humano necesita un tiempo de integración para reconocer el sonido que está escuchando, si bien es cierto que teóricamente este valor es menor a un segundo, se ha procedido con este valor como estándar por ser suficientemente representativo y válido para los cálculos necesarios.

Nótese que un segundo son 22050 muestras ya que la frecuencia de muestreo con las que han sido tomados cada uno de los ficheros es de $fs = 22050\, muestras/s$.

3.3.3.1 Energía y logEnergia por trama

La energía por trama se calcula sobre ventanas de 441 muestras (20ms), se desplazadas la mitad de su tamaño, 220 muestras (10ms). En nuestro caso, por cada fichero que se analiza, se obtienen 150 valores de la energía, ya que cada fichero tiene una duración de un segundo y medio

$(33075\ muestras_{fichero}/220\ muestras_{1/2 ventana})$.

Teniendo en cuenta que se analizan 650 ficheros se consigue un vector final de 96850 valores, el cual permite realizar una estimación bastante buena del comportamiento de las clases a analizar.

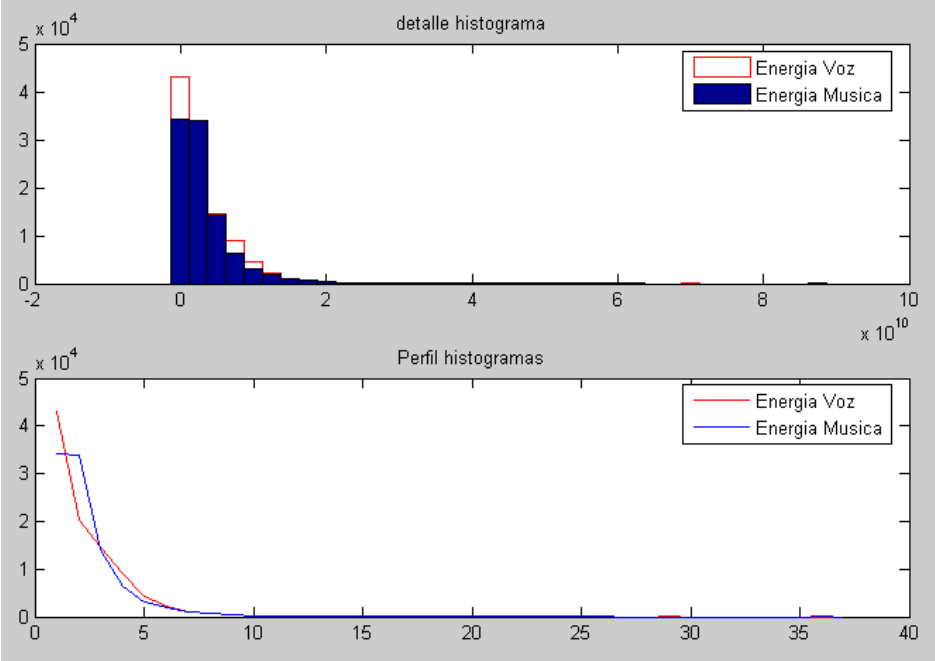

Figura 9: Histogramas y perfiles superpuestos de las energías de voz y música

En la *Figura 9* se representan los perfiles de los histogramas de la energía para cada clase calculados cada uno sobre 650 ficheros de la base de datos.

En este caso, los histogramas muestran un comportamiento bastante similar, si bien es cierto que hay una concentración de valores bastante alta entre 0 y $2x10^{10}$. Por tanto resulta de interés centrar la atención en esa parte de los valores:

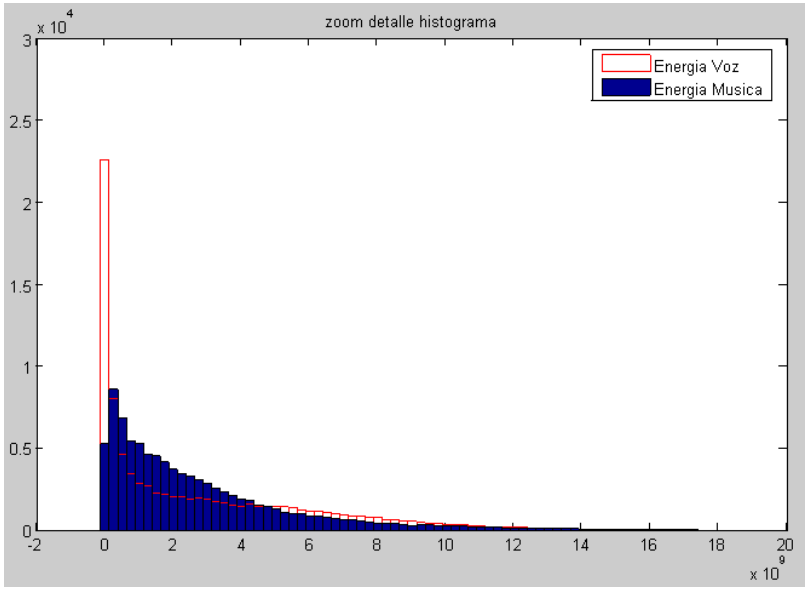

Figura 10: Detalle de los histogramas superpuestos de la Energía

Como se puede apreciar en la *Figura 10* la voz tiene una concentración muy alta de valores en las inmediaciones del cero, cerca de $2.5x10^4$ valores, mientras que la música, en ese mismo rango, maneja valores del orden de $0.5x10^4$.

Por otra parte, en valores superiores a $5x10^9$ la energía de la voz es superior a la de la música.

Como se intuyó en el apartado 3.3.2 de este mismo capítulo la voz cuenta con muchos valores próximos al cero, lo cual es debido a los silencios entre palabras y fonemas de energía baja, mientras que la música apenas tiene valores cercanos al cero ya que se comporta de una forma más regular y constante. Además, la energía de la música se agrupa de una forma clara en valores bajos, entre el 0 y el $4x10^9$, lo que indica que la intensidad de la música se mantiene en unos niveles más constantes. Ahora bien, si se analiza el histograma de la voz se puede apreciar una gran cantidad de valores próximos

al cero, como se ha indicado anteriormente, y una distribución de valores altos superior a los de la música, debido, principalmente, a las grandes variaciones sonoras que tiene la voz (*Figura 7*), que se compone de zonas de silencio y de zonas con gran cantidad de energía, como las vocales.

A partir de la energía por trama se calcula la logEnergia, que es simplemente realizar el logaritmo del vector de la energía. La logEnergía se utiliza para visualizar de una forma más clara las posibles diferencias entre las dos clases de archivos (voz y música) que se analizan.

El resultado gráfico de los histogramas de la logEnergía es el siguiente:

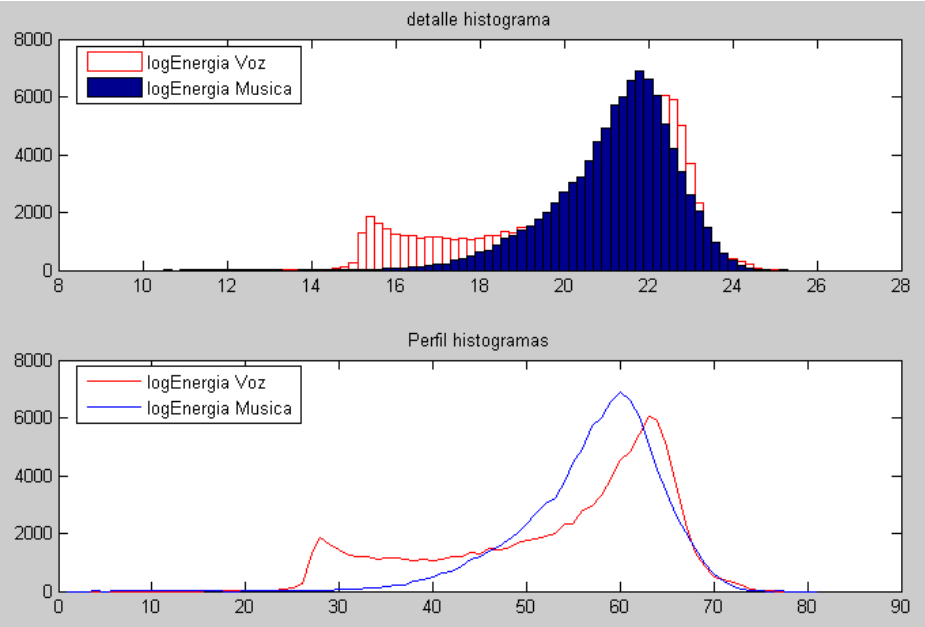

Figura 11: Histogramas y perfiles superpuestos de la logEnergía de voz y música

Como se aprecia en la representación de los histogramas ambas distribuciones se centran alrededor del 22. Asemejándose más a una campana de Gauss la música por tener una distribución más regular de valores, mientras que la voz, aún centrándose más o menos en el mismo punto que la música posee a lo que podemos hacer referencia como un lóbulo secundario centrado en valores inferiores y que reflejan en cierta medida los valores próximos a cero que se comentaron con referencia a la *figura 9 y 10*.

3.3.3.2 Media de la Energía y logEnergía en segmentos de un segundo

A continuación se analiza la media de la energía en segmentos de un segundo.

Las gráficas obtenidas para este tipo de análisis son las siguientes:

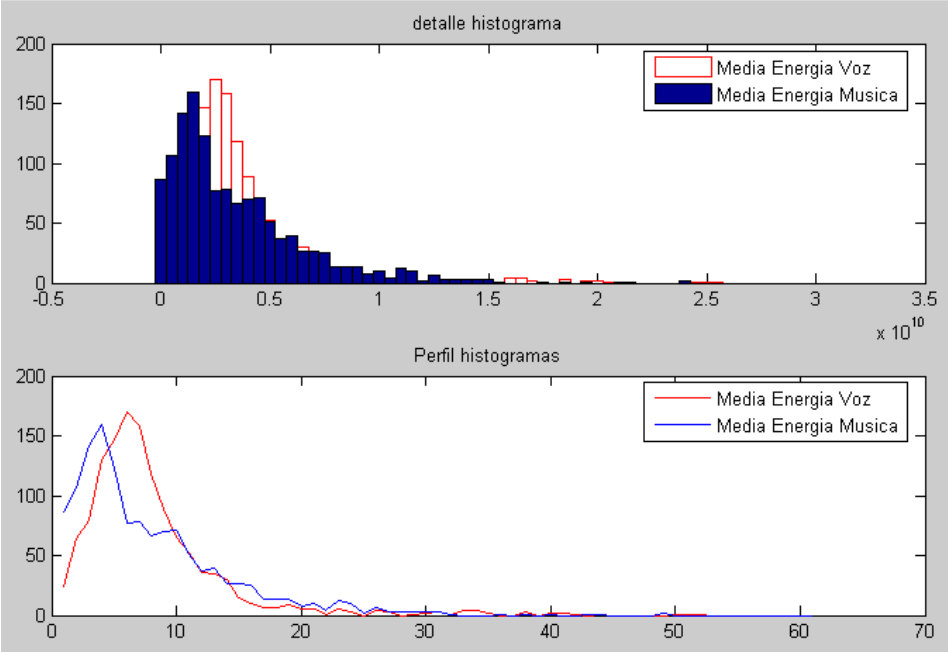

Figura 12: Histogramas y perfiles superpuestos de la Media de la Energía de voz y música

Como se observa en el histograma de la *Figura 12* y de una forma más clara en el perfil de la misma figura la distribución de la Media de la Energía se centra en valores inferiores a los de la voz. Esto puede llamar la atención en un primer momento ya que parece que es el resultado contrario a lo concluido en el apartado anterior *3.3.3.1*, pero si se dedica un poco más de tiempo a analizar el porqué de está situación la respuesta aparece de forma clara. Si bien es cierto que la voz tiene más tramas con energía próxima a cero, no es menos cierto que tiene más tramas con energía mayor que las de la voz (recordar lo explicado en referencia a la *figura 10*), así pues y por tanto al ponderar en segmentos de un segundo la energía media de la voz es superior que la de la música. Es decir, en la voz se promedian en un segundo valores próximos a cero

con valores muy altos, mientras que en la música el promedio se realiza de valores más o menos constantes y en media más bajos que los de la voz.

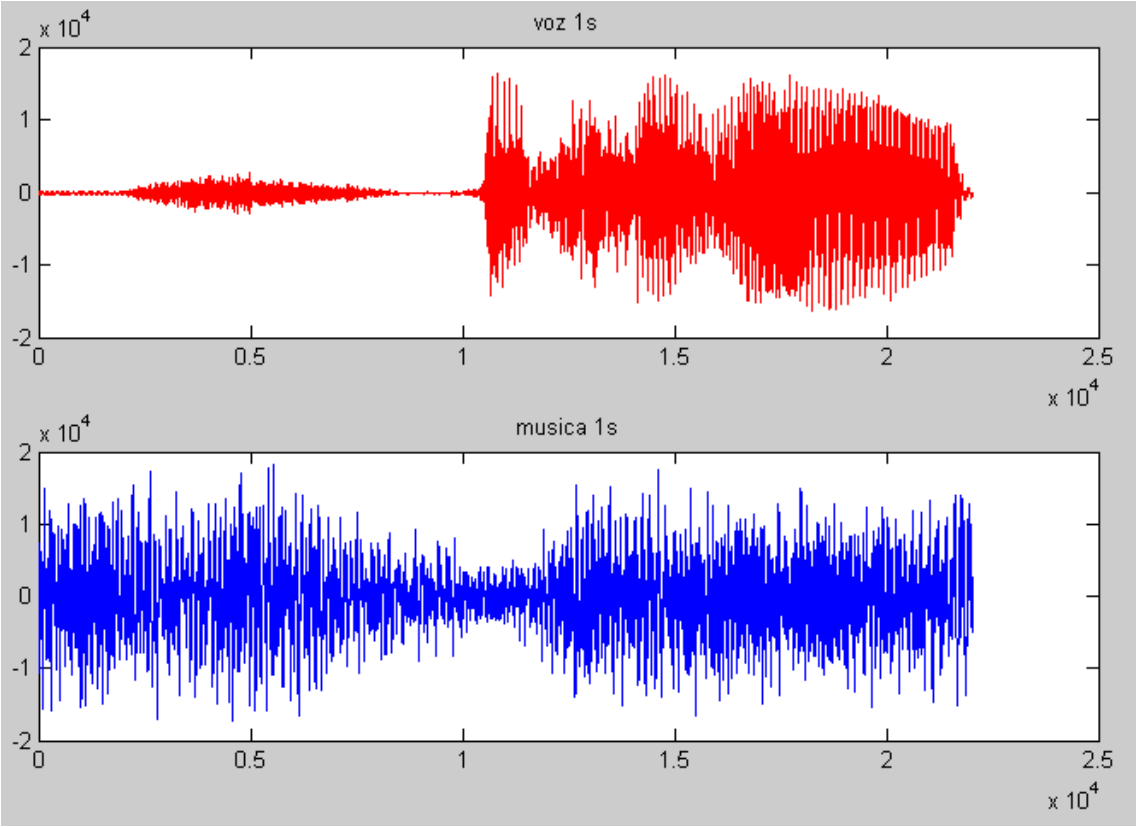

Figura 13: Representación de un segundo de voz (rojo) y uno de música (azul)

En la *figura 13* se aprecia un segundo al azar de una muestra de voz y de una de música. Se puede observar las grandes variaciones que sufre la muestra de voz con una parte inicial casi nula y una parte final de gran energía, mientras que el segmento de un segundo de música se mantiene muncho más constante. Así pues y debido a esto sucede que en media haya segmentos de voz con más energía que los segmentos de música.

La media de la logEnergía se lleva a cabo buscando una representación gráfica que sea más clarificadora a la hora de poder decidir lo que se parecen o no las distribuciones de las medias de las energías.

El histograma que se obtiene al realizar la media del logaritmo de la energía es el siguiente:

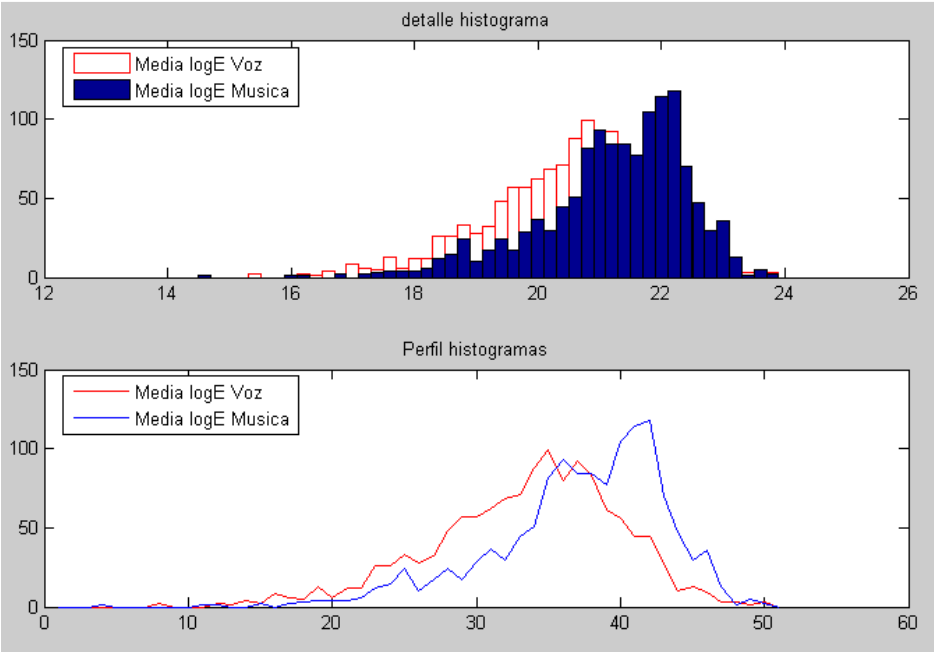

Figura 14: Histograma superpuestos de las medias de la logEnergía

El histograma de las medias de la logEnergía es muy similar al que se obtenía con el logaritmo de la Energía por trama, si bien el de la *figura 14* muestra un aspecto más irregular, sobre todo en el caso de la música.

Del mismo modo, y a pesar de las irregularidades, los histogramas que se analizan en este punto presentan distribuciones bastante similares. La música centrada en un valor ligeramente superior pero ambas con perfiles similares a campanas de Gauss.

3.3.3.3 Varianza de la Energía y logEnergía en segmentos de un segundo

Otro de los parámetros que analizamos en este proyecto es la varianza de la energía y del logaritmo de la energía para segmentos de un segundo.

Es cierto que las medias de ambos parámetros (energía y logE) son bastante parecidas, con algún punto de diferencia claro pero en líneas generales muy similares. Debido a ello se analiza la varianza de cada una de ellas, para ver como se ajusta este otro tipo de desviación.

El histograma de la varianza de la energía que se obtiene es el siguiente:

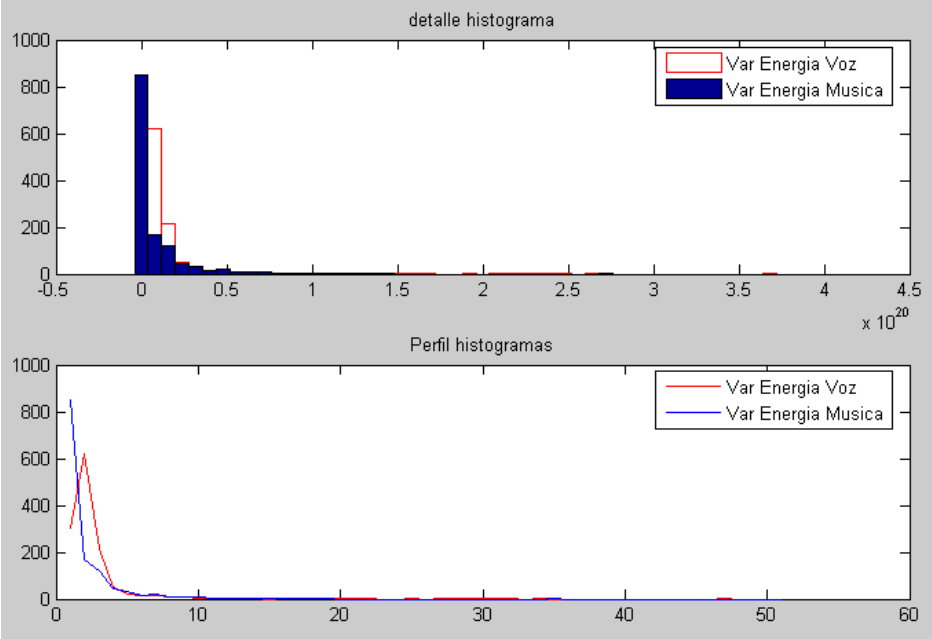

Figura 15: Histogramas superpuestos de las varianzas de la Energía

En la figura 15 puede observarse una gran concentración de valores en los parámetros próximos a cero y mucha dispersión hasta valores próximos de $4x10^{20}$. Si bien es cierto que si nos fijamos en el perfil de los histogramas aparece algo que llama la atención, la varianza de la música se concentra en cero mientras que la varianza de la voz lo hace en valores ligeramente superiores.

Analicemos esto con más detalle:

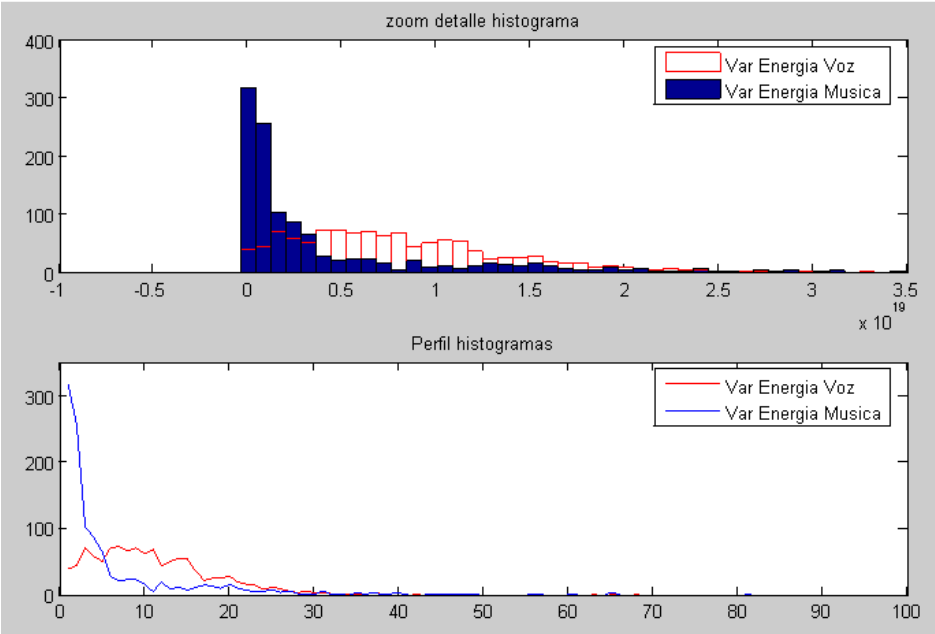

Figura 16: Detalle histogramas superpuestos de las varianzas de la Energía

En esta vista al detalle de las varianzas de las energías se aprecia más claramente como se distribuyen dichas varianzas. Así pues y como era de esperar la varianza de la música se concentra en valores próximos a cero, ya que como se ha comentado anteriormente la música tiene un comportamiento más regular. Mientras que la varianza de la voz se distribuye por valores superiores a cero, lo que indica que la energía de la voz es muy diferente a lo largo del tiempo, sufre grandes variaciones, es decir, tiene una gran varianza.

En el caso del logaritmo de la varianza de la energía se obtiene la siguiente gráfica:

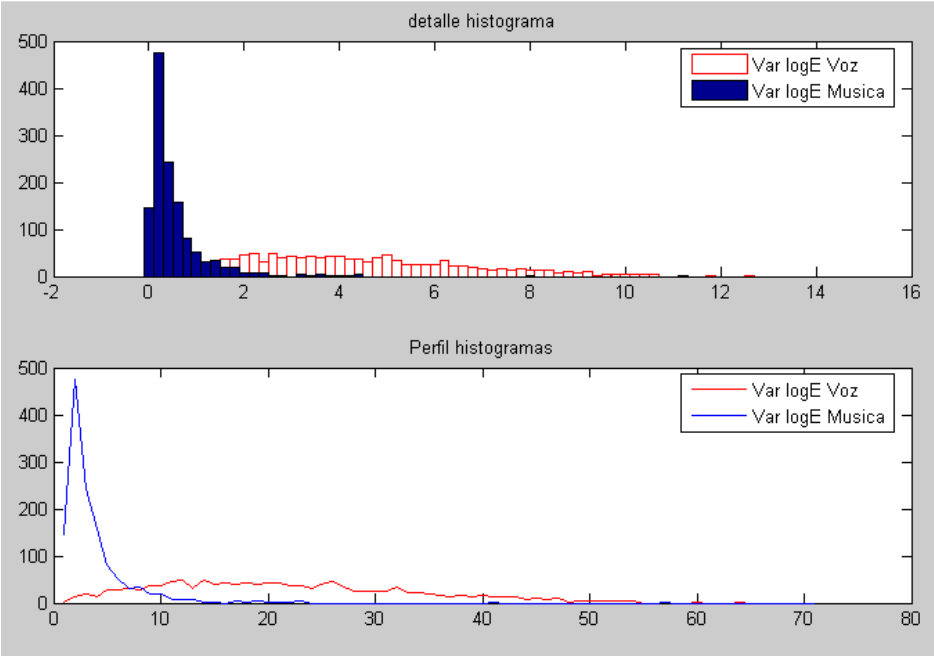

Figura 17: Histogramas superpuestos de las varianzas del logaritmo de la Energía

La varianza del logaritmo de la energía tiene una distribución muy parecida a la de la varianza de la energía, con la única diferencia en la escala en la que se engloban los valores, ya que en el caso de la varianza de la energía los valores oscilan entre 0 y $4x10^{20}$, y en la varianza de la logEnergía lo hacen entre valores de 0 y 14.

Se puede concluir, por tanto, que estos dos tipos de distribuciones son muy diferentes entre sí. Concentrándose ambos en valores bajos, pero de una forma mucho más pronunciada y relevante en el caso de la música, y de una manera más regular en los valores de la voz.

3.3.3.4 Porcentaje de tramas con energía baja

Finalmente se va a analizar el parámetro correspondiente al porcentaje de tramas de la señal de audio que tienen energía baja. Se entenderá por energía baja todas aquellas tramas con valores de la energía inferiores al 50% de la energía media del segmento.

El resultado que se espera es que la música tenga un porcentaje menor de tramas con energía baja, mientras que la voz tendrá una mayor cantidad de tramas que no llegan al umbral marcado de *Energía Baja*. Esto se debe a lo explicado en los apartados anteriores.

El resultado gráfico obtenido en el análisis del porcentaje de tramas con energía baja es el siguiente:

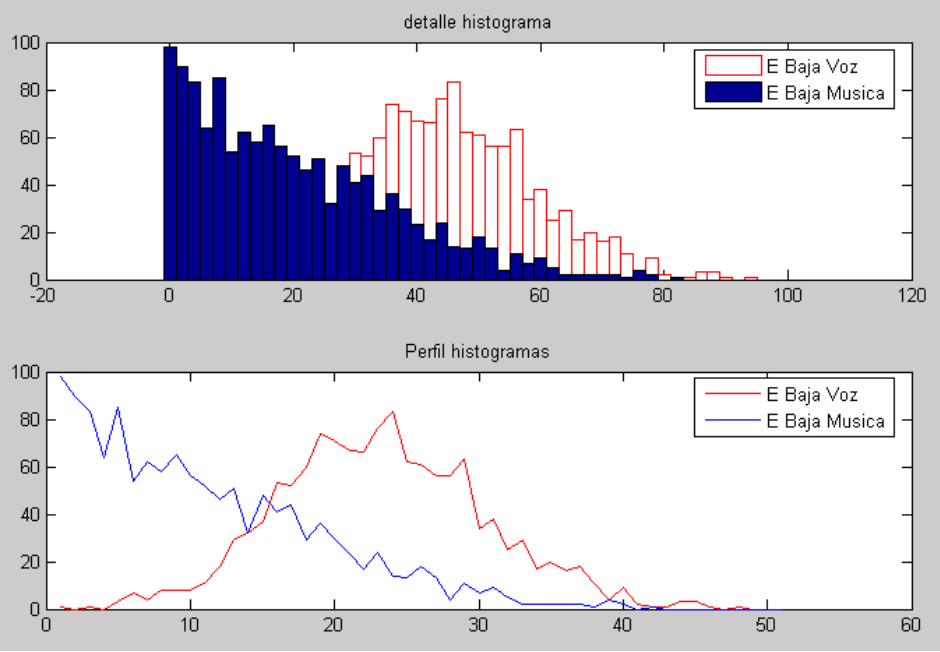

Figura 18: Histogramas superpuestos del porcentaje de tramas con energía baja

El comportamiento más regular de la música a lo largo del tiempo hace que la mayor parte de las tramas tengan una cantidad de energía suficiente que es superior al umbral marcado como energía mínima. Por el contrario, el comportamiento irregular de la voz lleva a que un gran número de tramas sean de silencio o con un valor muy bajo y por tanto no alcancen el umbral.

Como se puede observar en la *Figura 18* las distribuciones de ambas muestras (música y voz) presentan un aspecto de comportamiento muy diferenciado. Con una tendencia clara por parte de la música a concentrarse en valores bajos (poco porcentaje de tramas con energía baja) y una marcada disposición por parte de la voz a concentrarse en valores medios, con porcentajes del 40% ó 50% de las tramas que no superan el umbral de energía mínima marcado.

3.3.4 Distancia KLs entre distribuciones de energía

En este apartado se va a representar de forma cuantitativa lo estudiado gráficamente con los histogramas del apartado 3.3.3 (*Histogramas energía*). Así pues, se da un valor numérico a la semejanza o diferencia entre las dos distribuciones analizadas (recordar el apartado 3.2.3 *Distancia KLs*). Hay que tener en cuenta que los valores próximos a cero indican que dos distribuciones son muy similares, mientras que según va aumentando el valor refleja una menor semejanza entre parámetros analizados. Hay que tener en cuenta que la distancia de Kullback-Leibler no está acotada entre valores de 0 y 1, si no que toma valores desde 0 en adelante.

3.3.4.1 Distancia KLs para la energía por trama

Las dos distribuciones de la energía por trama tienen un comportamiento bastante diferente, pero hay que considerar que ese comportamiento tan diferente se lleva a cabo en una escala de 10^9, por lo que las diferencias no serán tan grandes.

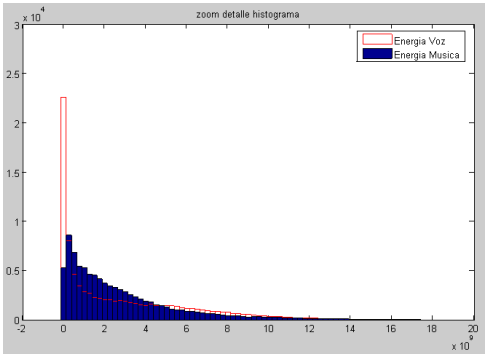

Figura 19: Histogramas Energía

El valor de Distancia KLs obtenido en este caso es de **0.0158**, valor bastante bajo y próximo a cero, por lo que aún teniendo comportamiento bastante diferente, en el conjunto general los histogramas se parecen bastante. Esto se subsanará con la realización del logaritmo.

Con la realización del logaritmo la escala disminuye a valores, en este caso, entre 10 y 28, por lo que el valor KLs obtenido será mayor que en el caso anterior.

Distancia KLs logEnergia = ***0.4016***

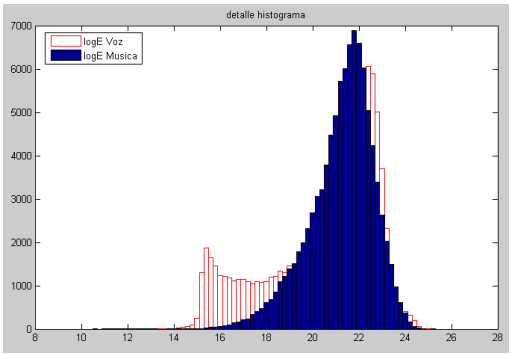

Figura 20: Histogramas logEnergía

3.3.4.2 Distancia KLs media de la energía por segmento

La energía media en segmentos de un segundo tiene un comportamiento bastante similar en la música y en la voz y eso es lo que nos refleja un valor de Kullback-Leibler de **0.0035**.

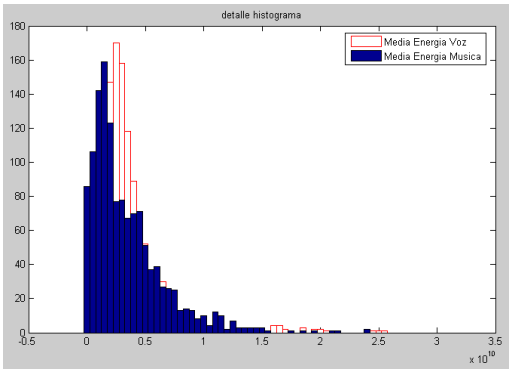

Figura 21: Histogramas Energía Media

En el caso de la logEnergia media por segmento se obtienen dos gráficas que aún teniendo un comportamiento bastante parecido ofrecen un valor cuantitativo de la distancia de Kullback-Leibler mayor que el caso de la energía media por segmentos.

DistanciaKLs = ***0.1756***

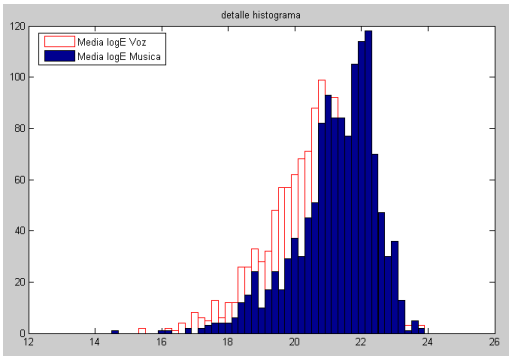

Figura 22: Histogramas logEnergía Media

3.3.4.3 Distancia KLs varianza de la energía por segmento

En el caso de la varianza en segmentos de un segundo el comportamiento es claramente diferente tanto para la energía como para el logaritmo de la energía. Ambos comportamientos son similares pero con la diferencia, una vez más y como es lógico, de la escala en la que ambas distribuciones se reparten, siendo del orden de 10^{19} en el caso de la varianza de la energía y del orden de decenas en el caso de la varianza del logaritmo de la energía.

KLs Var Energía = ***0.3632***

KLs Var logEnergia = ***9.0706***

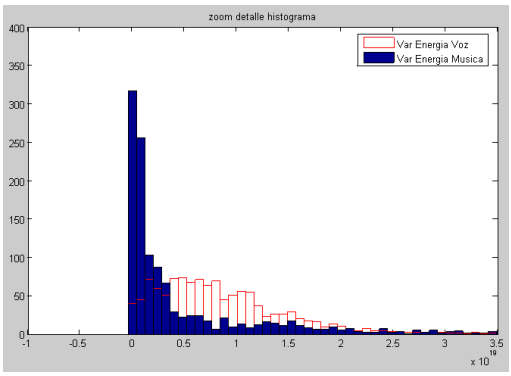

Figura 23: Histogramas Varianza Energía

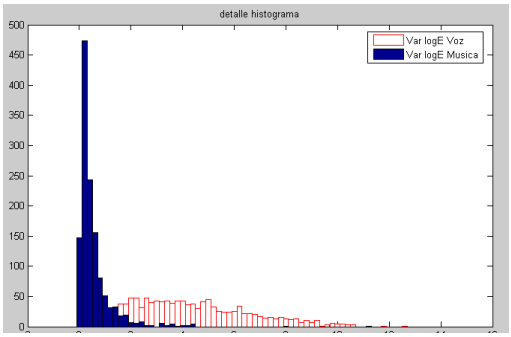

Figura 24: Histogramas Varianza logEnergia

3.3.4.4 Distancia KLs porcentaje de tramas con energía baja

Como se observa gráficamente las distribuciones de porcentajes de tramas con energía baja tienen un comportamiento claramente diferenciado, y la distancia KLs simplemente va a ser reflejo de esa diferencia de comportamiento.

El valor de la distancia de Kullback-Leibler para este caso es de **1.3004**

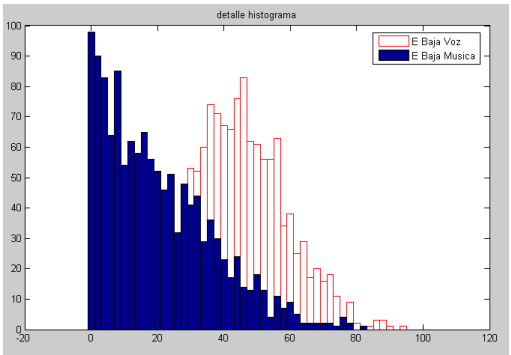

Figura 25: Histogramas % Energía Baja

3.4 Tasa de Cruces por Cero (ZCR)

Se denomina *cruce por cero* al hecho de que muestras consecutivas tengan diferente signo algebraico, es decir, el *cruce por cero* se considera cuando la señal a analizar pasa de positivo a negativo o viceversa. Por tanto y con esta idea de partida, se considera *tasa de cruces por cero* (ZCR Zero Crossing Rate) la cantidad de veces que una señal cambia de signo en una cantidad de tiempo determinada. [3.5]

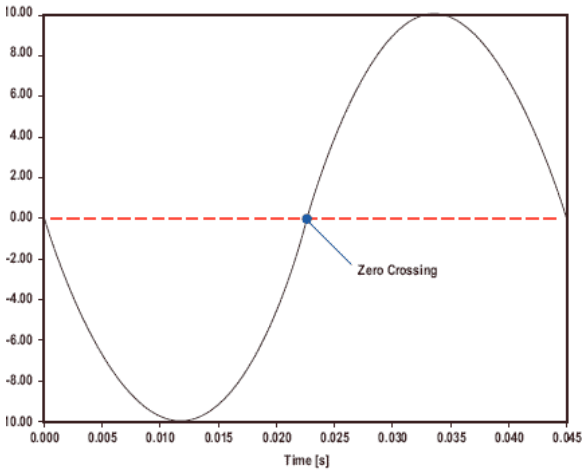

Figura 26: Cruce por cero

3.4.1 Cálculo de la Tasa de Cruces por Cero

Como en el parámetro analizado en el apartado anterior, para poder realizar una medición la señal a procesar se ha de limitar temporalmente, para lo cual se escoge de nuevo la ventana tipo H*amming*.

La tasa de cruces por cero localizada se define matemáticamente como [3.6]:

$$Z_s(m) = {}^{1}/_{N} \sum_{n=m-N+1}^{m} \frac{|sign\{s(n)\} - sign\{s(n-1)\}|}{2} w(n-m)$$

Ecuación 9: Fórmula de la Tasa de Cruces por Cero

Donde *w(n)* representa la ventana *Hamming*, N es la longitud de la ventana y la función signo toma los valores:

$$sign\{s(n)\} = \begin{cases} +1, s(n) \geq 0 \\ -1, s(n) < 0 \end{cases}$$

Ecuación 10: Fórmula de la función signo

Hay que tener en cuenta que aunque la ZCR es un parámetro calculado en el dominio del tiempo puede proporcionar información del contenido frecuencial de la señal.

La tasa de cruces por cero nos da una idea del carácter sordo o sonoro de la señal, entendiendo que el carácter sordo va ligado a tramo de alta frecuencia y por tanto tendrá una tasa de cruces por cero mayor.

Como ejemplo, se muestra a continuación los valores de ZCR para un segmento sonoro y otro sordo:

Segmentos sonoros:

- Energía por debajo de 3kHz
- Componente principal ± 700Hz
- ZCR = 14 en 10 ms

Segmentos sordos:

- Energía por encima de 3kHz
- Componente principal ± 2.5kHz
- ZCR = 49 en 10 ms

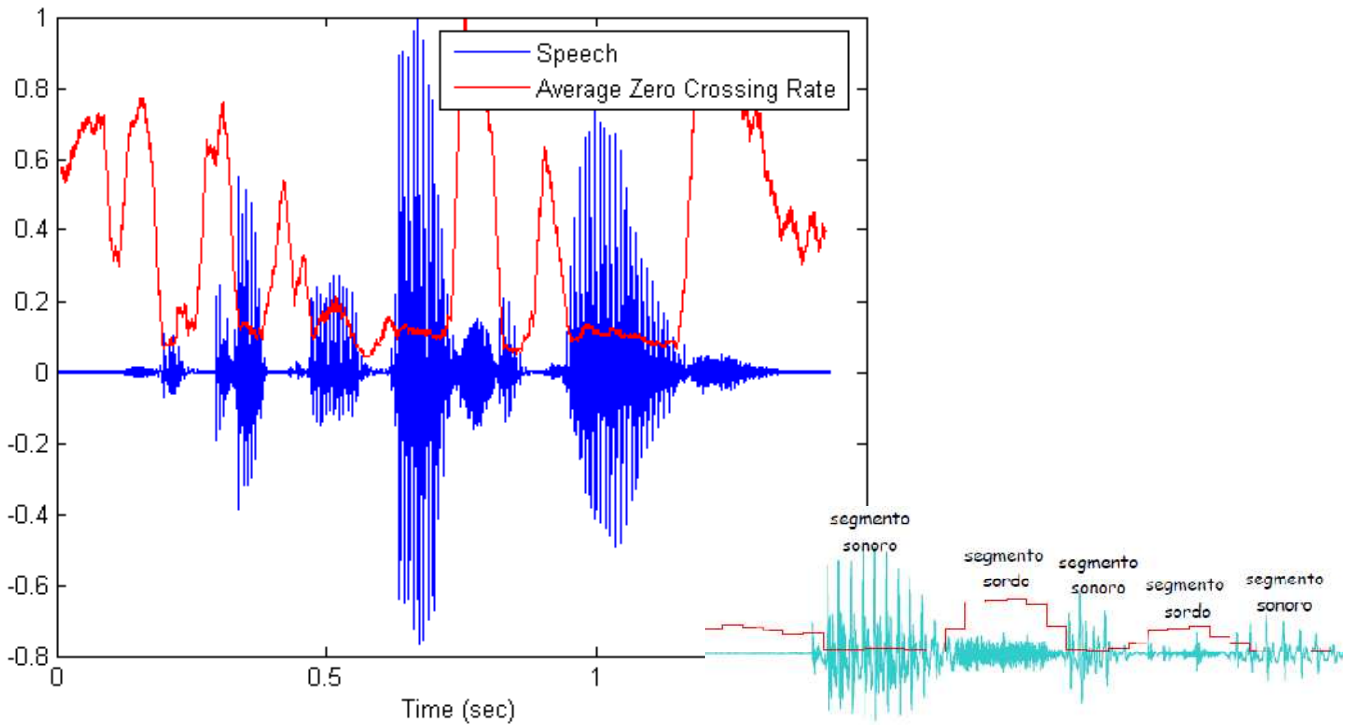

Figura 27: Comparativa visual elementos sordos y sonoros

En la *figura 27* [3.7] se puede ver gráficamente lo expuesto anteriormente. Tramos sonoros con una frecuencia menor a los tramos sordos, lo que hace que estos últimos fluctúen entorno al valor de cero, y por tanto tengan una tasa de cruces por cero mayor.

3.4.2 Idea intuitiva sobre la Tasa de Cruces por Cero

Una vez explicado lo anterior es complicado tener una idea clara de partida.

Lo que se espera con la tasa de cruces por cero es que la voz tenga una tasa mayor que la música, ya que posee una mayor cantidad de tramos sordos. Como se ha explicado anteriormente, los tramos sordos cuentan con una mayor cantidad de cruces por cero.

Si bien es cierto que la cantidad de tramos sordos y sonoros de la señal de voz puede llevar a que la ponderación media de la tasa de cruces por cero sea similar a la de la música. Ahora bien, teóricamente se espera que la música, al

estar compuesta principalmente de tramos sonoros, tenga una ZCR menor a la voz.

También hay que tener en cuenta, que aunque la música se componga en su mayor parte de tramos sonoros tiene sonidos que se comportan como sordos, teniendo en esos tramos una tasa de cruces por cero alta.

3.4.3 Histogramas de la Tasa de Cruces por Cero (ZCR)

En este apartado se van a mostrar los histogramas obtenidos en el análisis gráfico de la Tasa de Cruces por Cero por trama y la media y la varianza de la ZCR en segmentos de un segundo

3.4.3.1 Tasa de Cruces por Cero por trama

La tasa de cruces por cero por trama indica de forma cuantitativa entre 0 y 1 la cantidad de cruces por cero que tienen las muestras. El resultado es el siguiente:

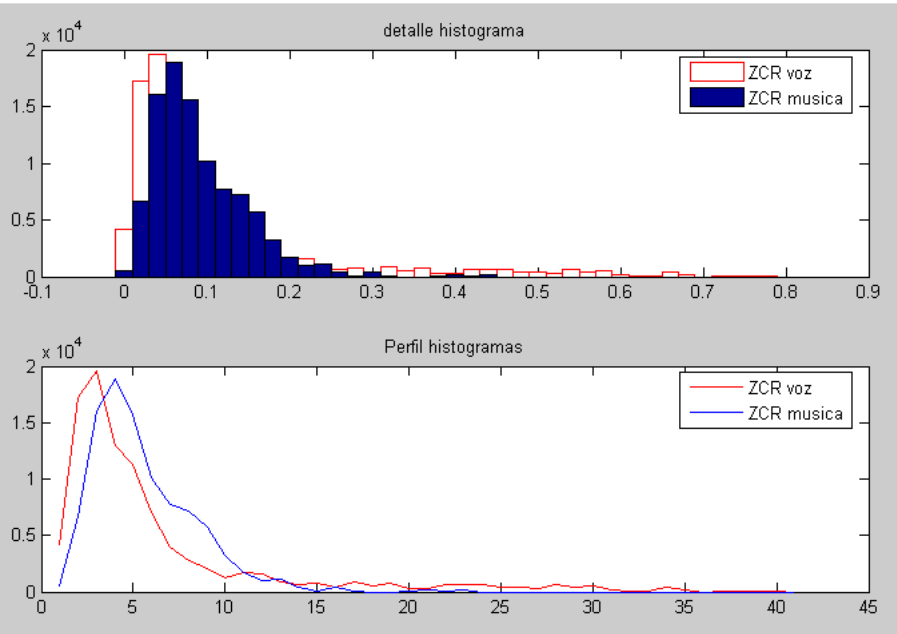

Figura 28: Histograma Tasa de Cruces por Cero

A pesar de lo que se podía esperar se obtiene una tasa de cruces por cero mayor para la música que para la voz, aunque hay que considerar que ambas distribuciones se asemejan bastante.

Y a pesar de que la distribución de la música se centra para un valor superior al de la voz (en torno a 0.08 frente a 0.05 para la voz) para valores superiores a 0.3 en la tasa de cruces por cero es la voz la que tiene estas cantidades más altas.

3.4.3.2 Media de la Tasa de Cruces por Cero en segmentos de un segundo

El resultado obtenido en este apartado es prácticamente igual al desarrollado en el anterior, con unos histogramas sorprendentemente similares:

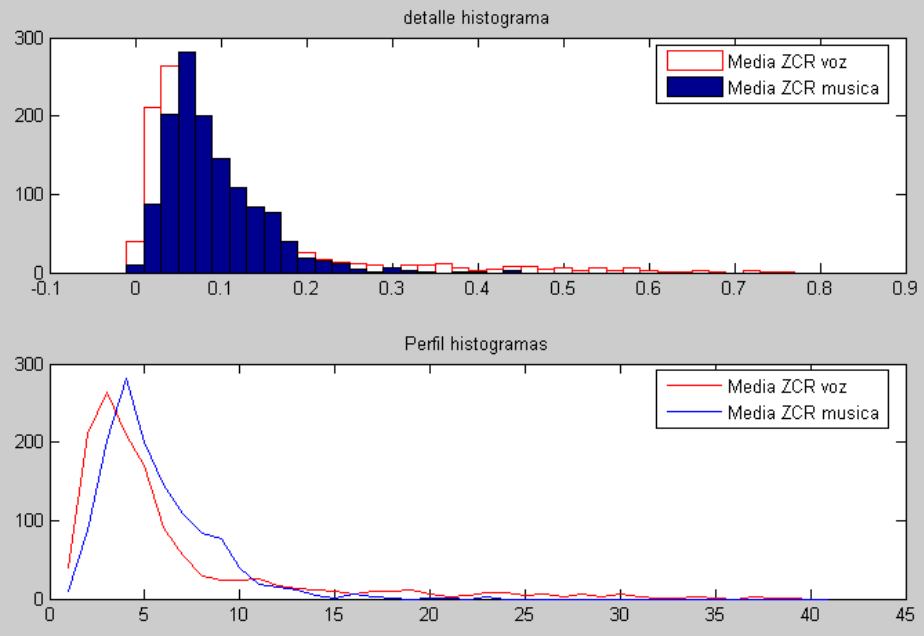

Figura 29: Histograma Tasa de Cruces por Cero Media

Las distribuciones que se obtienen son prácticamente iguales que en el apartado anterior, aunque en este caso los valores son calculados por el valor medio de la ZCR en segmentos de un segundo de duración.

3.4.3.3 Varianza de la Tasa de Cruces por Cero en segmentos de un segundo

Para este caso las dos distribuciones de la varianza también son prácticamente idénticas, con valores de varianza que se concentran sobremanera en el cero y sus proximidades.

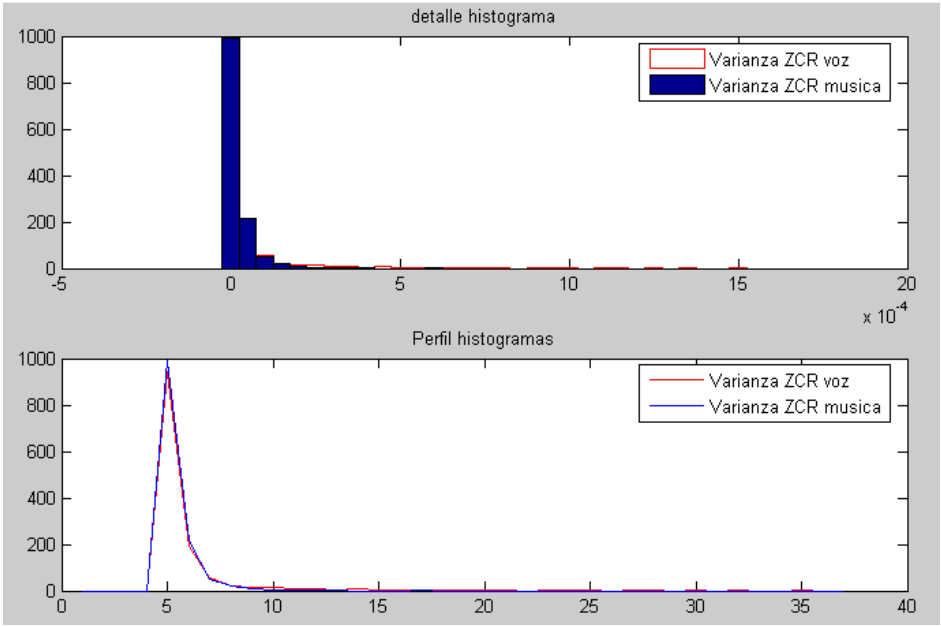

Figura 30: Histograma Tasa de Cruces por Cero Media

Así pues, en este caso, el análisis gráfico apenas nos da información, ya que en los tres casos analizados para la tasa de cruces por cero los histogramas de la voz y de la música son prácticamente idénticos.

3.4.4 Distancia KLs de la Tasa de Cruces por Cero

En este apartado se va a dar un valor numérico que indica la diferencia entre el histograma de la música y de la voz para cada uno de las gráficas presentadas en el apartado anterior.

Con este valor de la distancia de Kullback-Leibler se espera conseguir información adicional a la que ofrecían las gráficas, con la intención de poder usar estos parámetros para discernir entre música y voz.

3.4.4.1 Distancia KLs de la Tasa de Cruces por Cero por trama

A pesar de la semejanza gráfica de las dos distribuciones tienen una distancia de Kullback-Leibler suficientemente notable: **0.9572**

Esto es principalmente debido a que en un pequeño margen de valores (entre 0 y 0.2) se concentran la mayor parte de la tasa de cruces por cero, por tanto, cualquier mínima diferencia es realmente notable.

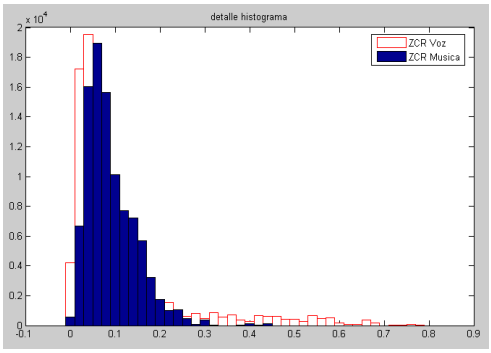

Figura 31: Histogramas ZCR

3.4.4.2 Distancia KLs de la Media de la Tasa de Cruces por Cero en segmentos

Como en el caso anterior, ambos histogramas tienen distribuciones claramente parecidas, por lo que gráficamente es muy complicado llegar a alguna conclusión.

Así pues, la distancia KLs nos ofrece un valor algo menor que en el apartado anterior, pero suficientemente representativo: **0.7755**

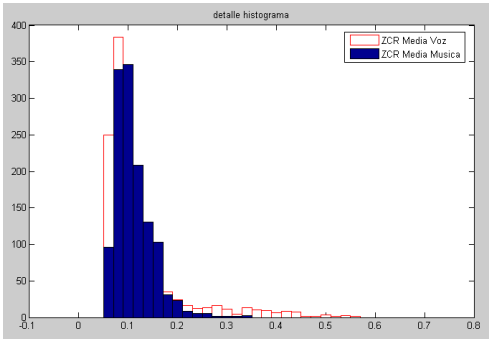

Figura 32: Histogramas Media ZCR

3.4.4.3 Distancia KLs de la Varianza de la Tasa de Cruces por Cero en segmentos

Como en los dos casos anteriores los histogramas siguen una distribución muy parecida tanto para la música como para la voz.

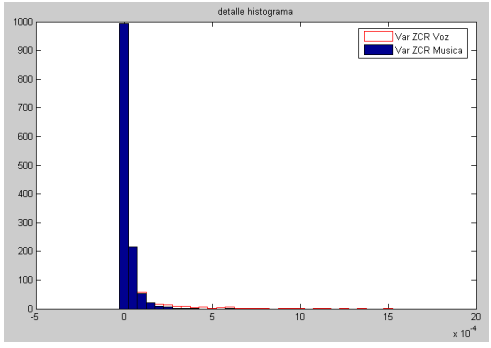

Figura 33: Histogramas Var ZCR

Pero como en el caso de la energía, el parámetro varianza es el que nos ofrece una distancia de Kullback-Leibler mayor, en este caso con un valor de **2.5011**. Si bien es cierto que no llega a los valores de la logEnergía (recordar que para ese caso la distancia KLs era del orden de 9) alcanza unos valores que son realmente significativos y habrá que tenerlos en cuenta.

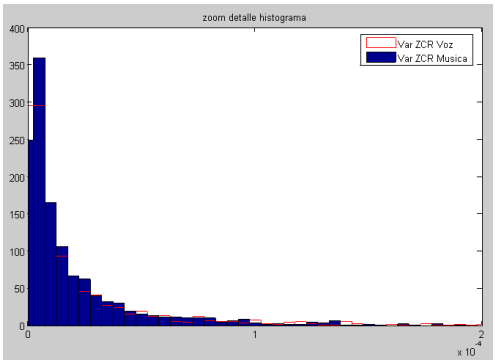

Figura 35: Valores Var ZCR próximos a cero

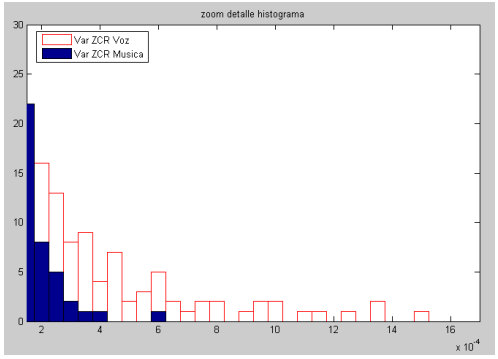

Figura 34: Valores Var ZCR alejados de cero

Las *figuras 34* y *35* nos muestran el detalle de los histogramas superpuestos de los valores de la varianza para la tasa de cruces por cero.

La *figura 34* representa los valores de la varianza de cero y próximos a cero, lugar donde se concentran gran cantidad de valores y donde apenas hay diferencia entre el histograma de la música y el de la voz.

La *figura 35* expone los valores más alejados de cero, y es aquí donde gráficamente podemos ver diferencias significativas entre ambos histogramas. Estas diferencias serán las que hacen que la distancia KLs sea significativa.

3.5 Frecuencia Fundamental (F0)

El último conjunto de parámetros considerado está relacionado con la Frecuencia Fundamental (F0) que se define como la frecuencia más baja del espectro de frecuencias tal que las frecuencias dominantes pueden expresarse como múltiplos de dicha frecuencia fundamental.

Un ejemplo bastante representativo es el siguiente: si se considera un objeto susceptible de vibrar como una cuerda de guitarra y analizamos las ondas emitidas por ésta, su frecuencia fundamental coincide con la frecuencia más baja en que esta cuerda puede vibrar estacionariamente. Cualquier sonido sostenido de esta cuerda podrá ser descompuesto como superposición de una vibración de frecuencia fundamental y armónicos superiores, es decir, vibraciones de frecuencias más altas que son múltiplos enteros de la frecuencia fundamental.

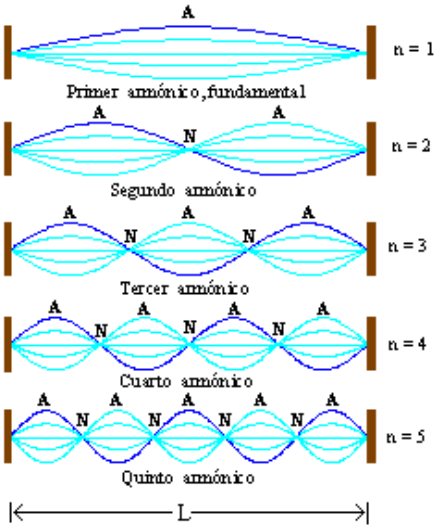

Figura 36: Frecuencia fundamental y sucesión de armónicos

En la *figura 36* se puede apreciar la frecuencia fundamental o primer armónico como la frecuencia más baja a la que la cuerda puede vibrar. Las otras cuatro figuras representan hasta el quinto armónico, múltiplos de la frecuencia fundamental.

Normalmente, a la hora de hacer vibrar un cuerpo, no obtenemos un sonido puro, sino un sonido compuesto de sonidos de diferentes frecuencias, estos son los armónicos. El sonido más próximo a una única frecuencia, a un tono puro, es el que se consigue con un diapasón, el resto de cuerpos ofrecen sonidos con múltiples armónicos.

3.5.1 Cálculo de la Frecuencia Fundamental

El cálculo de la frecuencia fundamental se realiza a partir de la toolbox de Matlab 'voicebox' [3.8], es un software libre que se puede utilizar y distribuir sin ningún tipo de restricción.

Dentro de esta toolbox de Matlab el algoritmo que se utiliza es el RAPT, un algoritmo robusto para el seguimiento o búsqueda del tono (Robust Algorithm for Pitch Tracking). Este algoritmo esta basado es la correlación cruzada y en la programación dinámica. No se utiliza solo la información local sobre la periodicidad, sino también la estimación de las tramas adyacentes a la búsqueda para una estimación global optima del tono fundamental. Así pues este es un complejo algoritmo que se basa en la Correlación Cruzada Normalizada (NCCF Normalized Cross-Correlation), en los transitorios de la voz y en la programación dinámica. [3.9]

En este apartado no solamente se analiza la frecuencia fundamental de la señal de voz y de la señal de música, si no que también se hará referencia a su media y varianza con sus respectivos logaritmos. Además, para tener una idea más completa de la información que aporta la frecuencia fundamental también se estudia el rango y el porcentaje de tramas sonoras.

Gracias al análisis de todas estas características se consigue una visión global más completa del comportamiento de la voz y de la música con respecto a la frecuencia fundamental.

3.5.2 Idea intuitiva sobre la Frecuencia Fundamental

No es fácil tener una idea clara del resultado que se puede obtener con respecto a la frecuencia fundamental, ya que el rango de la música es mucho mayor que el de la voz (puede cubrir el espectro de frecuencias completo). Si bien es cierto que el rango de la voz humana es mucho más limitado (hasta unos 10000Hz), suele tener la frecuencia fundamental en bajas frecuencias, menos de 200Hz; no hay que olvidar que a la hora del análisis frecuencial de la voz no se obtendrá un resultado igual si se trabaja con voces masculinas que con voces femeninas, teniendo estas últimas una frecuencia fundamental mayor y unos armónicos que también superan a los de la voz masculina, pudiendo llegar a 15000Hz.

Así pues, el resultado esperado de partida es que la frecuencia fundamental de la música sea más baja que la de la voz, pudiendo situarse la de la música entorno a los 50-75Hz de media frente a unos 150-200Hz por parte de la voz.

Por otra parte, y como se ha comentado al inicio de este subcapítulo, el rango de F0 de la música se espera mayor que el de la voz, cubriendo el primero casi todo el rango espectral y quedando el de la voz limitado sin llegar a muy bajas ni a muy altas frecuencias. No hay que olvidar que todo este análisis dependerá de cómo sean las muestras de música analizadas, ya que es posible que la señal de música esté muy restringida en frecuencias, aunque no es lo habitual y no es lo que se espera.

Por último, con respecto del análisis del porcentaje de tramas sonoras, se espera que este porcentaje sea mayor en la música que en la voz. Ya que como se comentó en el apartado anterior (ZCR) la voz cuenta con muchos segmentos sordos o de poca energía, mientras que la música tiene un comportamiento más regular. Así pues, el porcentaje de tramas sonoras en la música se puede esperar del 85-90%, mientras que el de la voz no superará en prácticamente ningún caso el 60%.

3.5.3 Histogramas Frecuencia Fundamental (F0)

En este apartado se va a analizar de forma gráfica cada una de las características procesadas de la frecuencia fundamental: F0 por trama y su logaritmo, media y varianza por segmentos y sus respectivos logaritmos; el rango de F0 y el porcentaje de tramas sonoras por segmentos.

3.5.3.1 Frecuencia Fundamental por Trama

El histograma de la frecuencia fundamental analizado por tramas muestra una tendencia más o menos clara de la música a situarse en frecuencias más bajas, entorno a los 75 Hz, mientras que la voz se centra en torno a los 100 Hz (lo que nos indica que las voces analizadas son principalmente masculinas).

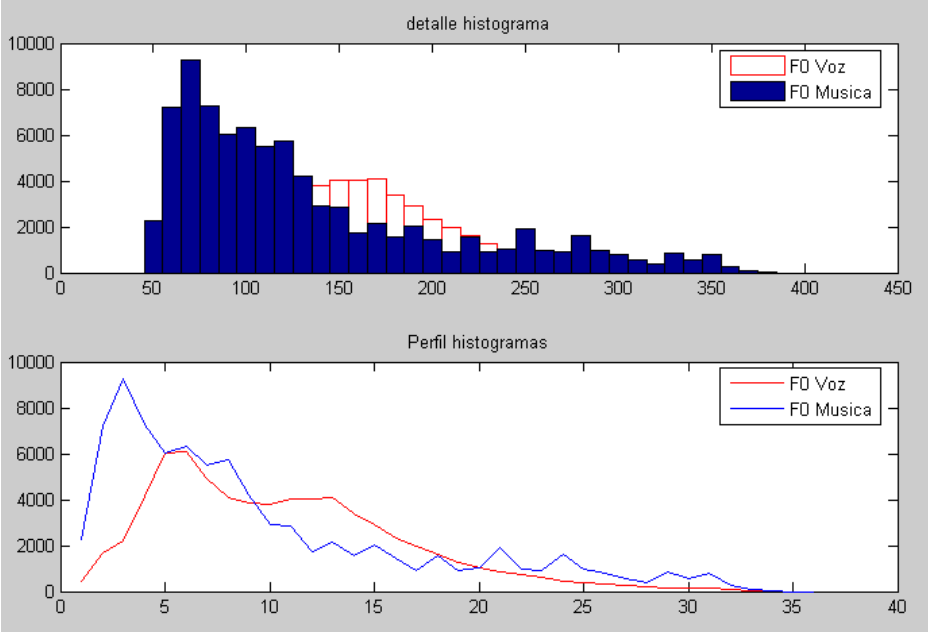

Figura 37: Frecuencia fundamental de la música (azul) y de la voz (rojo)

Si analizamos el histograma con un poco más de definición podemos observar lo siguiente (*figura 38*):

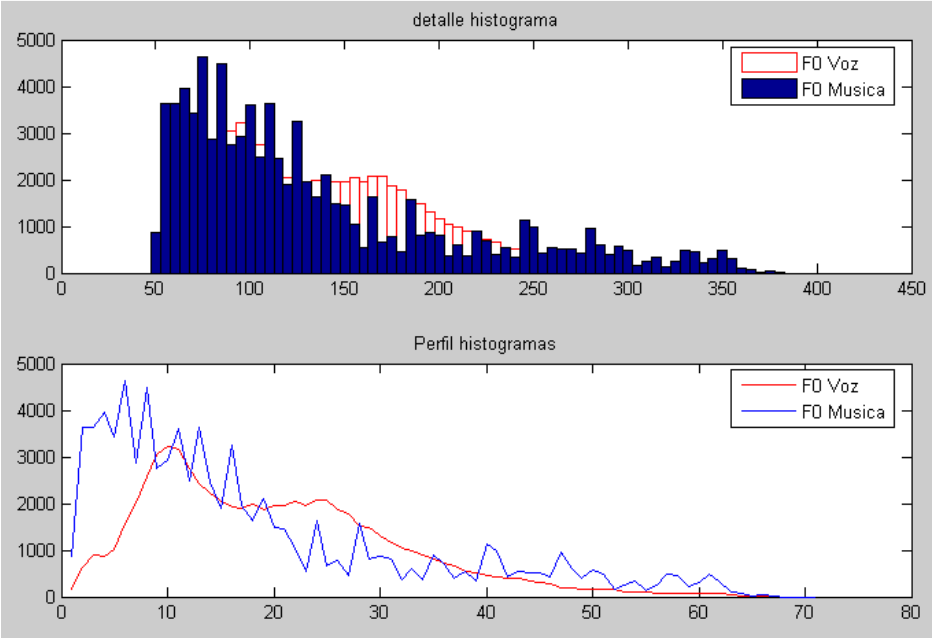

Figura 38: Frecuencia fundamental de la música (azul) y de la voz (rojo)

Si centramos la atención en el perfil de los histogramas se puede observar un comportamiento mucho más regular en la frecuencia fundamental de la voz (color rojo) frente al comportamiento 'crestado' de la música (color azul). Esto puede ser debido a que la frecuencia fundamental de la voz tiene unos valores que van a estar dentro de un margen más o menos delimitado en todas las tramas, mientras que la F0 de la música va a ser mucho más variante, debido a que, como ya se ha comentado anteriormente, cubre un espectro en frecuencias mucho mayor, lo que hace que dependiendo de cada trama la frecuencia fundamental pueda variar notablemente.

Si ahora se analiza el logaritmo de la frecuencia fundamental por trama se obtiene la siguiente gráfica:

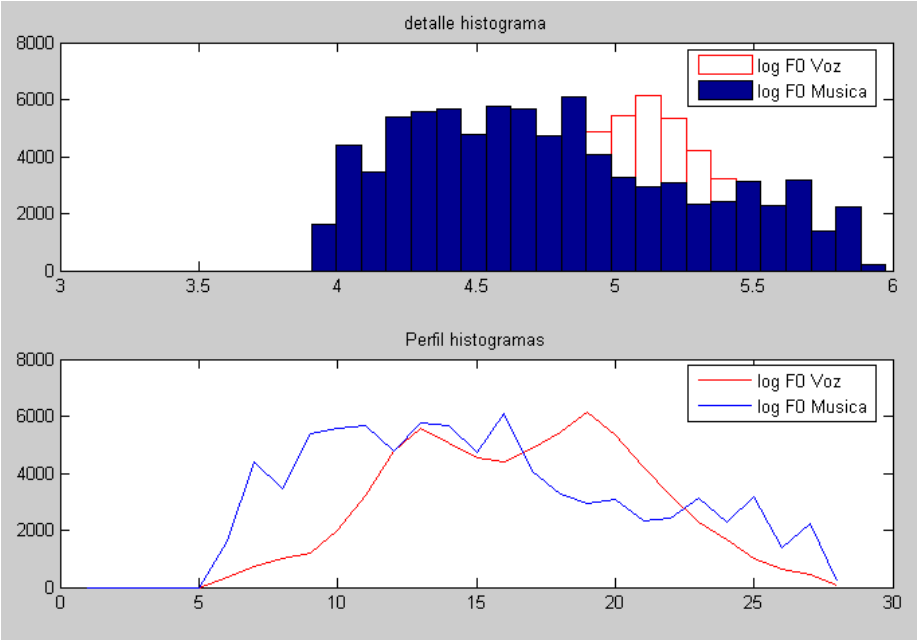

Figura 39: Logaritmo de la frecuencia fundamental de la música (azul) y de la voz (rojo)

Al realizar el logaritmo de valores relativamente bajos, del orden de las centenas, lo que se consigue al hacer el logaritmo es concentrar los valores en un rango muy pequeño, por lo que apenas obtenemos información extra al hacer el logaritmo.

Si fijamos la atención en el perfil del histograma se puede observar que las distribuciones del logaritmo de la frecuencia fundamental de la voz como de la música tienen un patrón muy similar. Debido a esto la información que se puede obtener de este parámetro parece más reducida.

3.5.3.2 Frecuencia Fundamental Media en segmentos de un segundo

A continuación se analiza el histograma obtenido para la media frecuencia fundamental calculada en segmentos de un segundo:

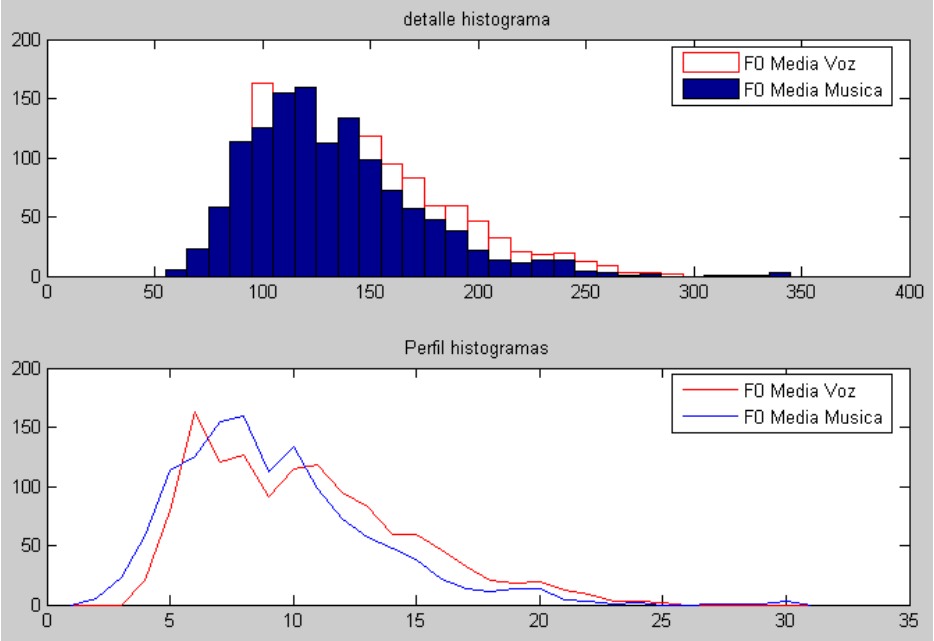

Figura 40: Frecuencia fundamental media en segmentos de un segundo

Como se observa en la *figura 40* el comportamiento de ambos histogramas es muy similar. Si bien es cierto que la frecuencia fundamental es menor en la música, cuando se realiza una ponderación media ambas frecuencias fundamentales se asemejan sobremanera, esto es debido a que la voz mantiene una F0 que se puede considerar más o menos constante en torno a 100Hz (*figuras 37* y *38*) mientras que en la música esta F0 varía considerablemente entre las diferentes tramas y cuando se realiza una ponderación se obtiene un valor medio similar al de la voz.

Si pasamos a analizar el logaritmo de ambos parámetros obtenemos unos histogramas aún más similares que los anteriores y que se asemejan de forma muy considerable a una campana de Gauss.

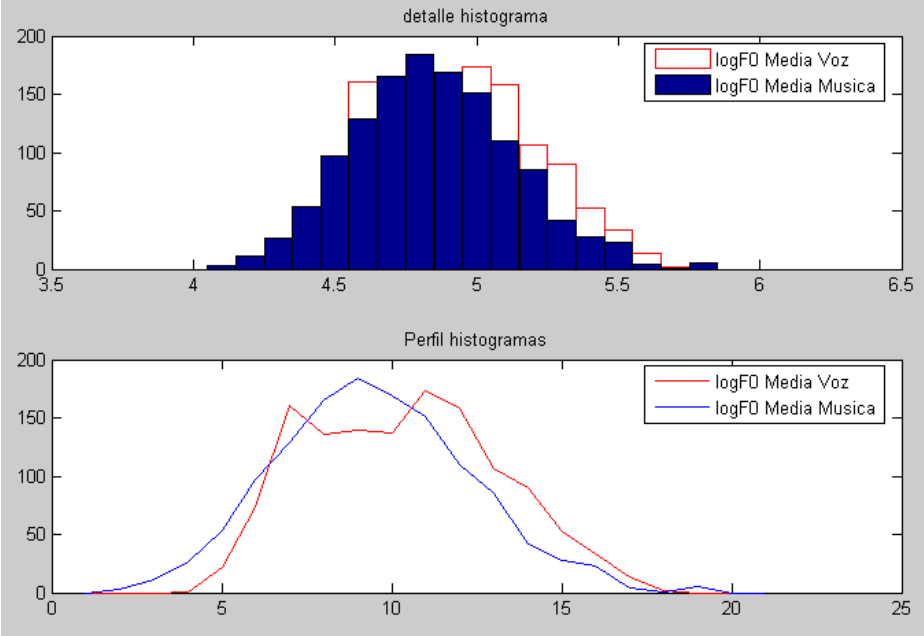

Figura 41: Logaritmo frecuencia fundamental media en segmentos de un segundo

Debido a estas claras similitudes la información que se puede obtener de estas gráficas es muy reducida, por lo que parece ser no será un buen parámetro para la discriminación entre música y voz.

3.5.3.3 Frecuencia Fundamental Varianza en segmentos de un segundo

La varianza de la frecuencia fundamental indicará la dispersión que sufre la F0 en la voz y en la música. De forma teórica este valor a de ser mayor en la música (cubre un rango espectral mayor) que en la voz.

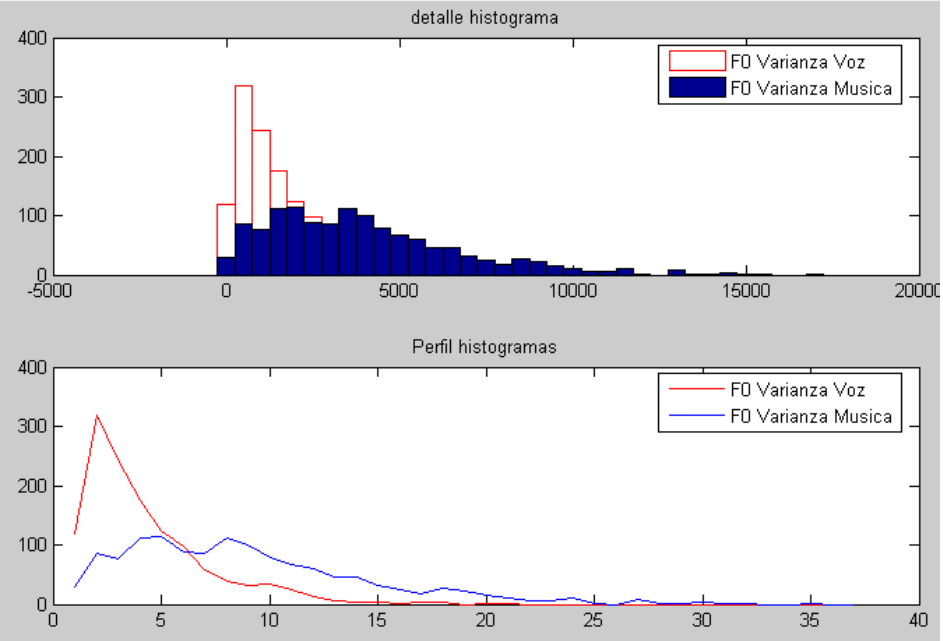

Figura 42: Varianza frecuencia fundamental en segmentos de un segundo

Como era previsible la distribución de la varianza de la frecuencia fundamental de la voz se concentra en valores bajos, mientras que la de la música lo hace de forma más dispersa, concentrándose de forma regular entre 0 y 5000 y sufriendo un claro descenso a partir de ese valor.

Si se analiza el logaritmo de la varianza se obtiene lo siguiente:

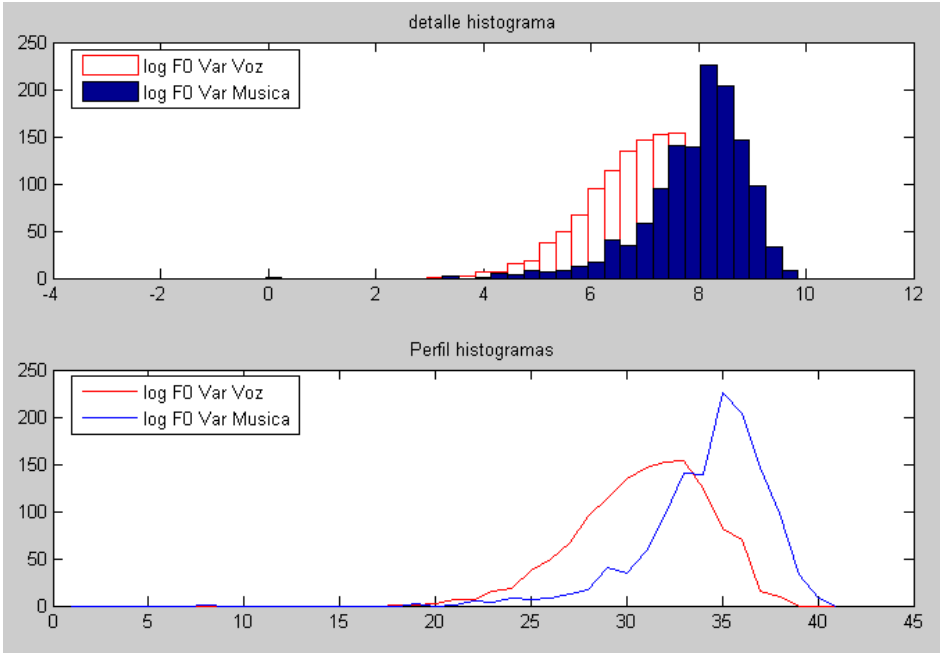

Figura 43: Logaritmo de la varianza de la frecuencia fundamental en segmentos de un segundo

Al realizar el logaritmo se logra una concentración de los valores, obteniendo dos gráficas que aún teniendo un comportamiento similar se concentran en valores que se pueden conseguir diferenciar, ya que la música tienen una tendencia a concentrarse en valores superiores a los de la voz.

3.5.3.4 Rango de la Frecuencia Fundamental en segmentos de un segundo

El rango de la frecuencia fundamental se calcula como la diferencia entre el valor máximo y el valor mínimo de frecuencia fundamental en cada segmento de un segundo.

$$Rango\ F_0 = Max\ (F_0) - \min(F_0)$$

Una vez realizado el procesado de la señal se obtiene la siguiente gráfica:

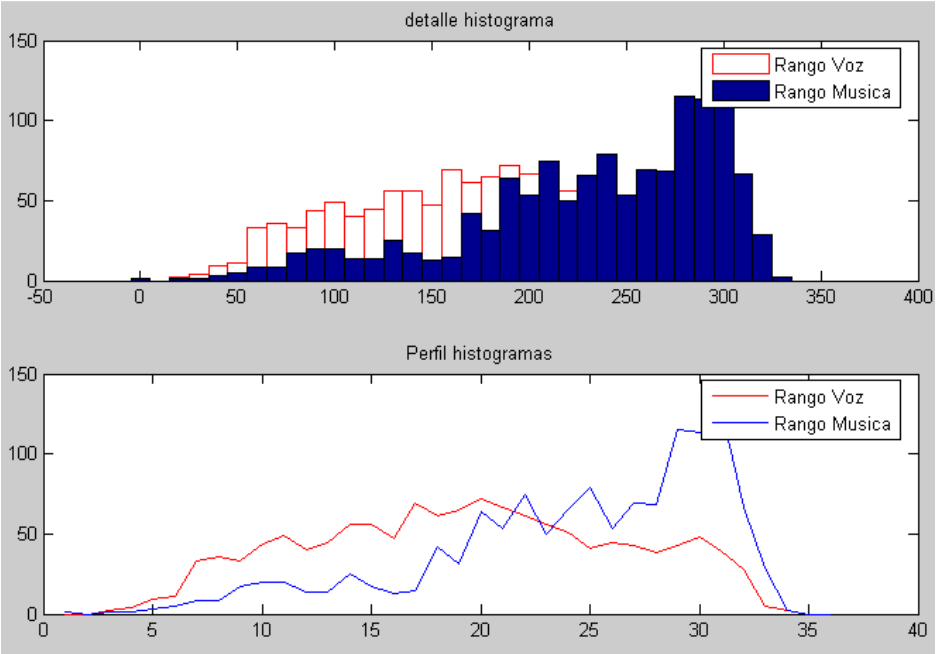

Figura 44: Rango de la frecuencia fundamental en segmentos de un segundo

Como se observa en el histograma de la *figura 44* el rango de la música se centra en valores superiores al de la voz. Esto es debido, como se ha explicado anteriormente, a que la señal de música cubre un rango espectral mayor, por lo que sus frecuencias fundamentales pueden variar dentro del mismo segundo de una forma más significativa que en el caso de la voz.

3.5.3.5 Porcentaje de tramas sonoras en segmentos de un segundo

Finalmente, se va a analizar el porcentaje de tramas sonoras, entendiendo por tramas sonoras aquellas que tienen un valor de frecuencia fundamental. Tanto en la música como en la voz (más en ésta) nos encontramos con tramas de la señal que no tienen F0, porque bien son silencio, bien son solamente ruido, o bien corresponden con fonemas sordos (en el caso de la voz).

Teniendo en cuenta esto, todas las tramas que no tienen un valor de F0 las consideraremos tramas sordas. Así pues, para el cálculo de este parámetro simplemente será necesario determinar el número de tramas sonoras.

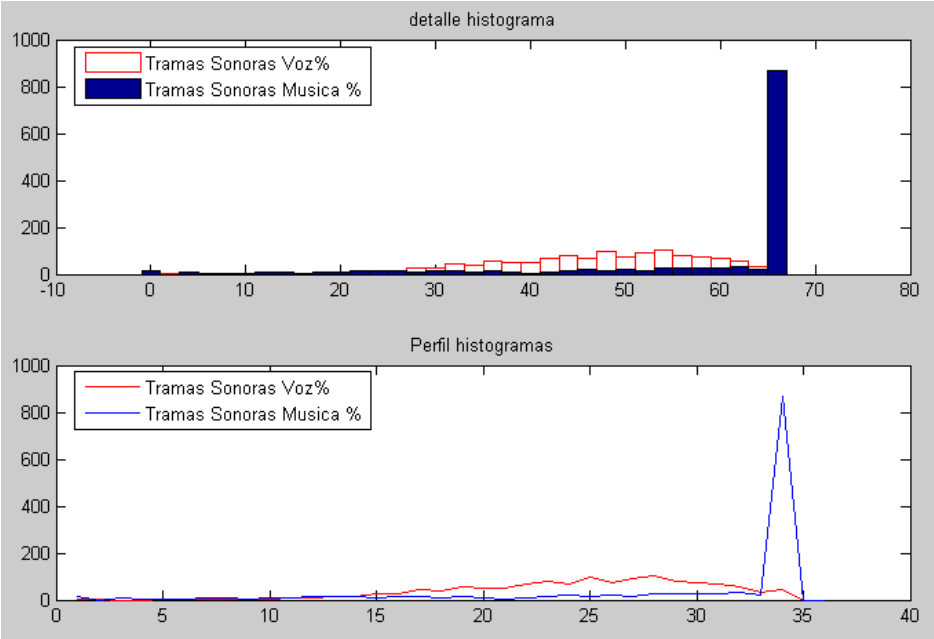

Figura 45: Porcentaje de tramas sonoras en segmentos de un segundo

Como se mostró en el *subapartado 3.5.2* y se observa en la *figura 45*, el porcentaje de tramas sonoras en la música es mucho mayor que en la voz.

En la música se centra en torno al 66%, 2 de cada 3 tramas son sonoras, mientras que en la voz es difícil saber con exactitud el punto central de la distribución, pero ronda el 50%, indicando que 1 de cada 2 tramas es sonora.

Las distribuciones obtenidas son claramente diferentes, por lo que dicho parámetro parece tener gran cantidad de información para la discriminación entre clases.

3.5.4 Distancia KLs de la Frecuencia Fundamental

Una vez estudiado el análisis gráfico de los histogramas para los diferentes parámetros de la frecuencia fundamental, en este apartado se intentará cuantificar la diferencia entre las distribuciones de la música y de la voz dentro de una misma característica.

3.5.4.1 Distancia KLs de la Frecuencia Fundamental por Trama

La distribución de la música se centra claramente en valores inferiores a los de la voz, esto conlleva a tener un valor de la distancia de Kullback-Leibler de **0.0838**. Valor bajo, ya que aunque las distribuciones teóricamente son claramente diferenciables en la práctica no se consigue una diferenciación tan definida.

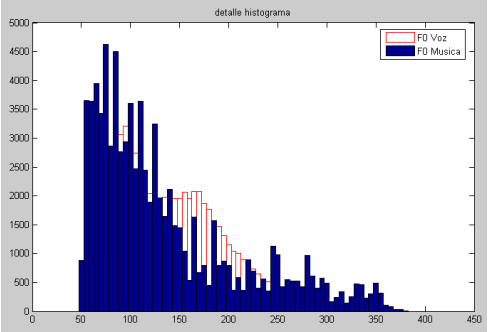

Figura 46: Histogramas F0

Al realizar el logaritmo de la frecuencia fundamental lo que se logra es una concentración de valores, por lo que la distancia KLs apenas varía: **0.1092**

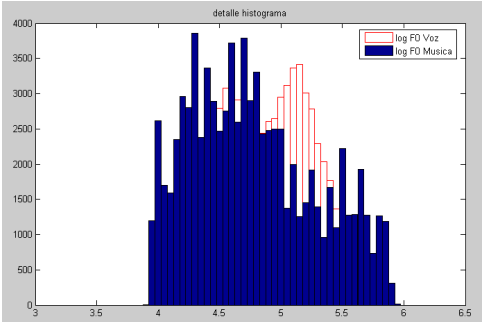

Figura 47: Histogramas log F0

3.5.4.2 Distancia KLs de la Media de F0 en segmentos de un segundo

Como se explicó en el apartado *3.5.3.2* los histogramas para estas dos distribuciones son realmente semejantes, obteniendo así una distribución KLs de **0.0419**

Obteniendo un valor prácticamente similar si se estudian las distribuciones de la media del logaritmo de F0: **0.0465**.

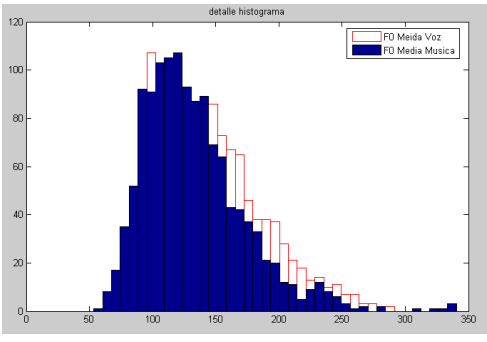

Figura 48: Histogramas Media F0

3.5.4.3 Distancia KLs de la Varianza de F0 en segmentos de un segundo

Como se aprecia en la figura de la derecha, la varianza de la voz se concentra en valores considerablemente más bajos a los de la música, las distribuciones son claramente diferentes, por lo que se obtiene un valor KLs de **1.2350**

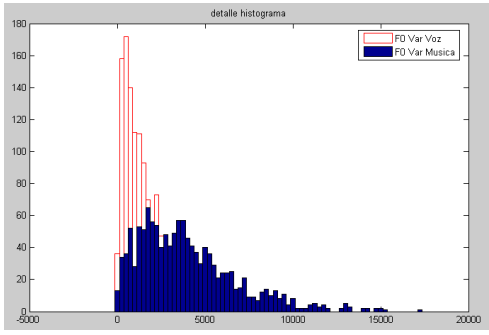

Figura 49: Histogramas Varianza F0

Si ahora se analiza la varianza del logaritmo se obtiene un valor menor, ya que las distribuciones de los logaritmos tienden a asemejarse.

KLs logVar = ***0.5200***

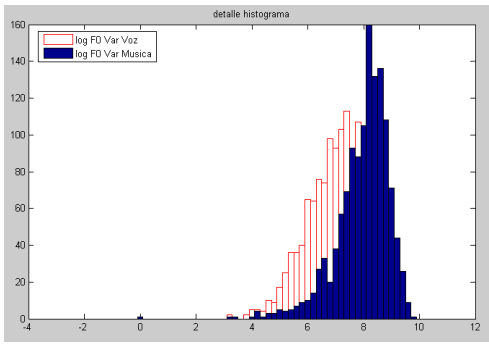

Figura 50: Histogramas logaritmo Varianza F0

3.5.4.4 Distancia KLs del Rango de F0 en segmentos de un segundo

Como era previsible el rango de la música se centra claramente en valores superiores a los de la voz. Si bien es cierto que el valor KLs no es todo lo grande que se espera debido a la gran superposición de valores entre ambas distribuciones.

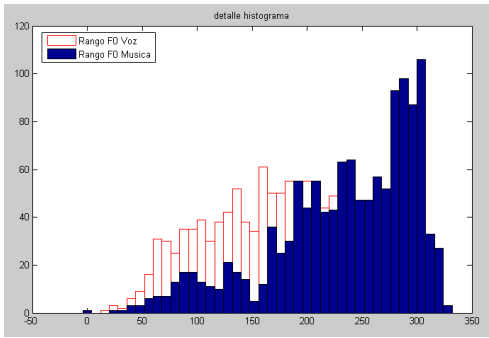

Figura 51: Histogramas Rango F0

KLs Rango F0 = ***0.2813***

3.5.4.5 Distancia KLs del Porcentaje de Tramas Sonoras en segmentos de un segundo

Como se aprecia claramente en la *figura 52* las distribuciones del porcentaje de tramas sonoras son realmente diferentes, debido sobre todo a que la distribución de la música se centra en un único valor de aproximadamente el 66%.

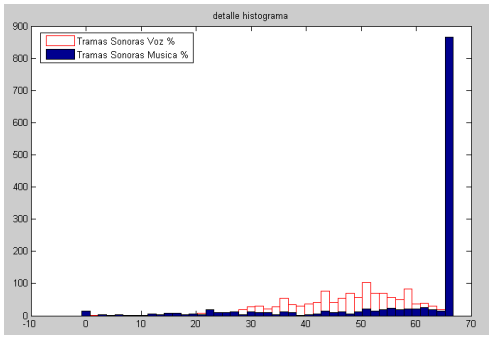

Figura 52: Histogramas Tramas Sonoras %

La distancia KLs que se obtiene puede sorprender por ser excesivamente baja, pero hay que tener en cuenta que ambas distribuciones se superponen durante un gran rango de valores.

Así pues se obtiene una distancia KLs: **0.4216**.

Capítulo 4. Sistemas de clasificación de registros de audio

4.1 Introducción

A lo largo de este capítulo se va a explicar los clasificadores utilizados para la categorización en dos clases de las pistas de audio, en ficheros de audio de música instrumental y en ficheros de audio de voz.

Un clasificador toma un conjunto de características como entradas y produce como salida una clase etiquetada. Una manera de construir un clasificador es mediante un conjunto de ejemplos etiquetados y tratar de definir una regla que pueda asignar una etiqueta a cualquier otro dato de entrada. [4.1]

Los clasificadores con los que se ha trabajado han sido el clasificador lineal y el clasificador cuadrático, en los cuales los datos de entrada son una característica acústica en concreto o la combinación de varias, de las características descritas en el capítulo anterior.

A lo largo de este capítulo se analiza el experimento de clasificación que se realiza para características analizadas por tramas y por segmentos (Capítulo 3). Las características analizadas y los resultados obtenidos se muestran en el capítulo 5 (Resultados Experimentales).

4.2 Procedimiento de clasificación

Para la clasificación se ha llevado a cabo uno de los procesos estándar en la clasificación de parámetros. Se han utilizado el 80% de las muestras para entrenar el clasificador (muestra 'Train') y el 20% de las muestras restantes para llevar a cabo la parte de test (muestras 'Test'). Este 20% de muestras es lo que realmente se clasifica y se verifica si el clasificador lo ha realizado correctamente, obteniendo de este modo un porcentaje de aciertos y de errores.

Se ha de tener en cuenta que de partida se sabe el resultado que el clasificador debe ofrecer, lo que permite comparar el resultado devuelto con los valores reales que se tienen de partida y así poder comprobar el porcentaje de acierto que ofrece el clasificador para las características que se estén analizando en cada momento.

Como el conjunto de las muestras puede considerarse que no es suficientemente representativo, se lleva a cabo lo que se conoce como validación cruzada.

4.2.1 Validación cruzada

Este método se lleva a cabo cuando el conjunto de muestras analizadas no llega a ser suficientemente representativo y se intenta buscar una mayor fiabilidad de los resultados obtenidos.

Para lograrlo lo que se lleva a cabo es lo que se puede denominar como una rotación de las muestras que componen el conjunto de análisis 'test'. Para ello se dividen todas las muestras con las que se trabajan en cinco grupos (en nuestro caso 160 ficheros por grupo (800/5)) y se realizan cinco clasificaciones, en cada una de ellas el grupo de test es uno diferente y los cuatro grupos restantes se utilizan como conjunto de entrenamiento (Figura 1). Finalmente, el

resultado final de la clasificación se calcula como el promedio de los resultados parciales.

Con este procedimiento se consigue una simulación como si se estuviese trabajando con 5 veces más de muestras, por lo que el resultado global obtenido es más real y, por tanto, ofrece una visión más exacta del comportamiento de la clasificación de ese parámetro o conjunto de parámetros.

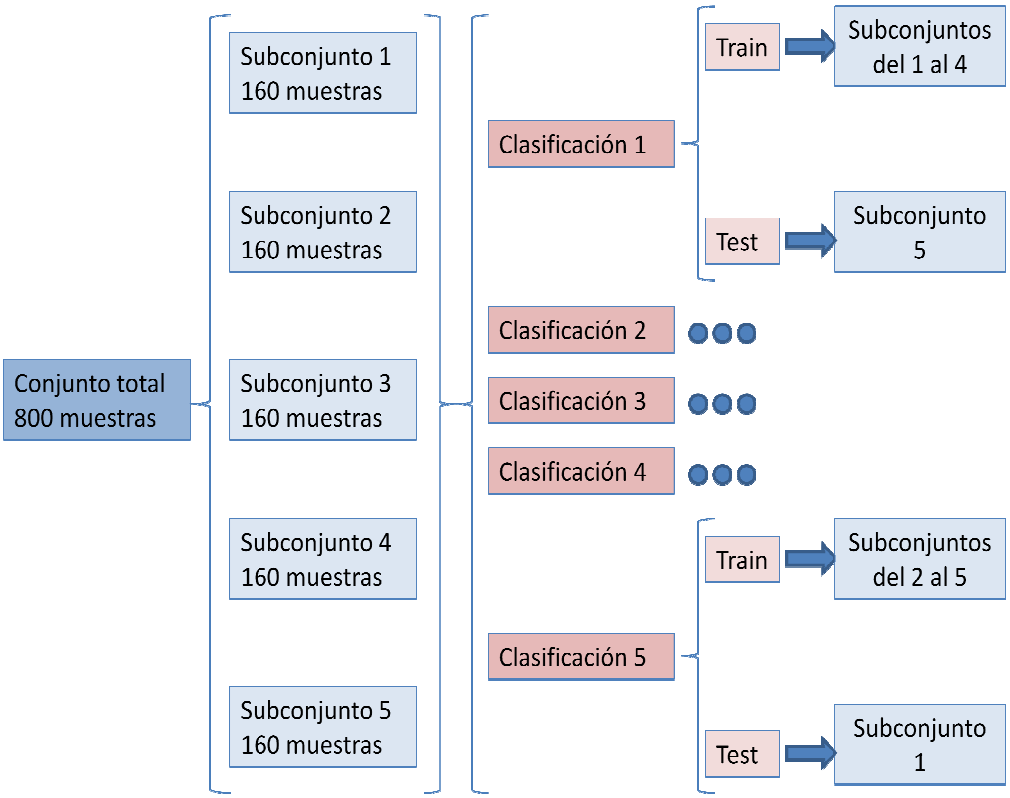

Figura 53 Validación Cruzada

4.3 Clasificador lineal

El clasificador lineal se caracteriza por crear una frontera de clasificación recta que divide el conjunto de muestras a clasificar en dos grupos definidos por esa frontera de clasificación, como se puede apreciar de forma didáctica en la figura siguiente:

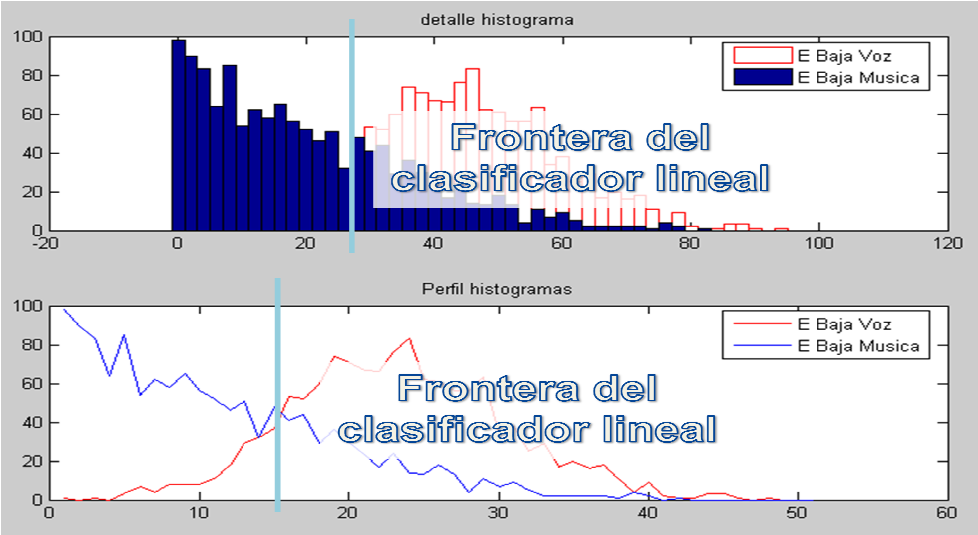

Figura 54 Ejemplo de frontera de clasificador lineal

Las fronteras de decisión se expresan como una función lineal.

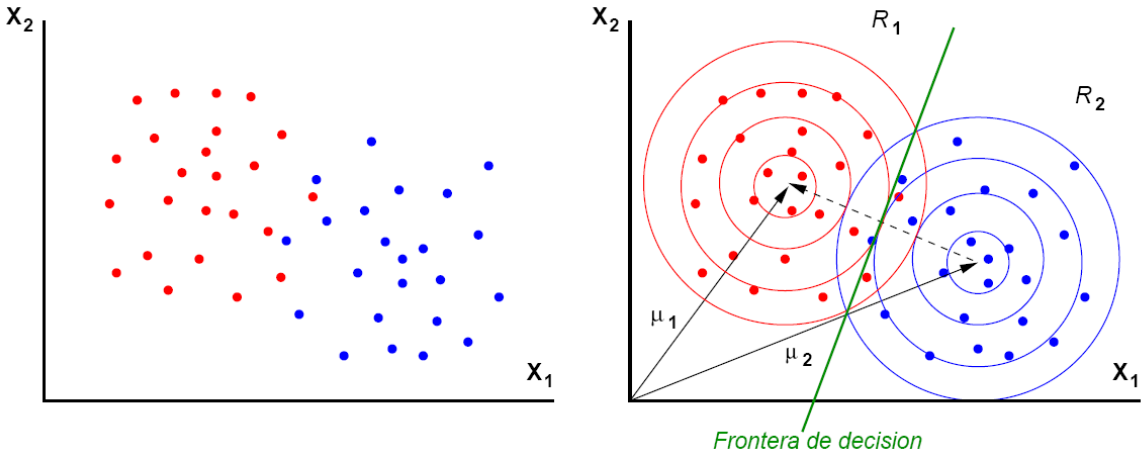

Figura 55 Centro de gravedad de las nubes y frontera de decisión lineal

Es lo que se considera un clasificador de mínima distancia. Se calcula la distancia euclídea de cada patrón a cada uno de los centros de las clases obtenidos en la parte de entrenamiento, y se asigna la etiqueta a la clase más cercana.

Es, quizá, el clasificador más simple con el que se puede trabajar, por lo que la el porcentaje de aciertos en la clasificación no siempre será el óptimo.

Por otra parte, hay que tener en cuenta, que si un clasificador lineal ofrece un porcentaje de acierto de elementos clasificados suficientemente alto, indicará que la característica o características tratadas en ese caso forman grupos o nubes claramente separables, lo que indicará un comportamiento disparejo de cada una de esas clases para esa característica en concreto.

4.4 Clasificador cuadrático

El clasificador cuadrático se caracteriza por crear una frontera de clasificación mediante combinaciones cuadráticas de las componentes del vector de características. Es decir, las fronteras de decisión se expresan como una función cuadrática (círculos, elipses, parábolas, hipérbolas).

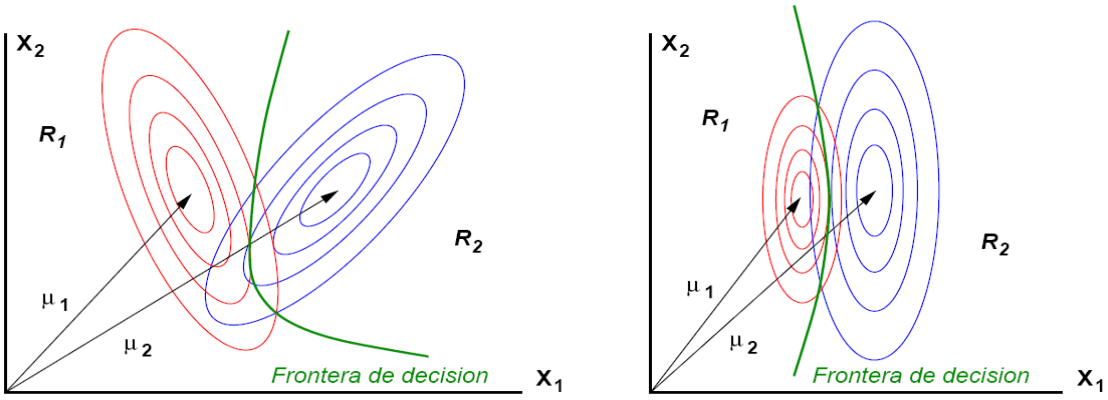

Figura 56 Centro de gravedad de las nubes y frontera de decisión cuadrática

Desde el punto de vista formal, el clasificador cuadrático es más general. Sin embargo, un clasificador cuadrático requiere un número de muestras mucho mayor que un clasificador lineal para estimar adecuadamente las densidades de probabilidad. Los estimadores no son fiables cuando tenemos pocos datos y/o la dimensionalidad de los datos es elevada.

Por tanto, aunque la decisión de adoptar un clasificador cuadrático o uno lineal depende fundamentalmente de la forma de las covarianza de las clases, el clasificador cuadrático requiere muchas más muestras de entrenamiento que un clasificador lineal para conseguir resultados similares.

Capítulo 5. Resultados experimentales

5.1 Introducción

En este capítulo se engloba todos los resultados obtenidos de forma experimental al aplicar sobre los clasificadores cada una de las características acústicas descritas en el capítulo 3 o diversas combinaciones de ellas.

Así pues y gracias al conjunto global de resultados obtenidos se podrá llegar a ciertas conclusiones sobre qué características ofrecen unos rendimientos de clasificación mayor y por tanto se pueden utilizar en líneas futuras de trabajo.

En este capítulo los resultados se presentan del siguiente modo:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	54.22	58.12	62.50	59.53	46.56	**56.17**
Cuadrático	50.78	60.62	56.72	49.84	53.44	**53.16**

Tabla 1 Ejemplo de cuadro de % de aciertos en relación al clasificador usado

Donde 'lineal' indica que se ha utilizado el clasificador lineal; 'cuadrático' hace referencia a que la clasificación se ha llevado a cabo con dicho clasificador.

Los valores referenciados en las casillas del grupo 1 al grupo 5 indican los aciertos de las cinco rotaciones de la validación cruzada con el clasificador lineal y del clasificador cuadrático. El valor más a la derecha en la tabla ('Media') indica los valores medios de la validación cruzada en cada uno de los casos.

5.2 Clasificación con parámetros trama a trama

En este apartado se describe el proceso experimental sobre las características obtenidas trama a trama de los ficheros de audio con los que se ha trabajado (*logEnergia, Tasa de Cruces por Cero y Frecuencia Fundamental*).

A continuación se muestran los experimentos realizados y los resultados obtenidos.

5.2.1 Resultados con la LogEnergia

Teniendo presente el histograma que se obtiene de distribución de valores de la logEnergía por trama (analizado en el capítulo 3 de este proyecto), el resultado esperado no puede ser muy superior al 50-60%, ya que ambas distribuciones se concentran en más o menos los mismos valores. Vamos a tener en cuenta también el valor de la distancia de Kullback-Leibler para ir definiendo a partir de que valores de esta distancia dos distribuciones se pueden empezar a diferenciar con cierto grado de aciertos.

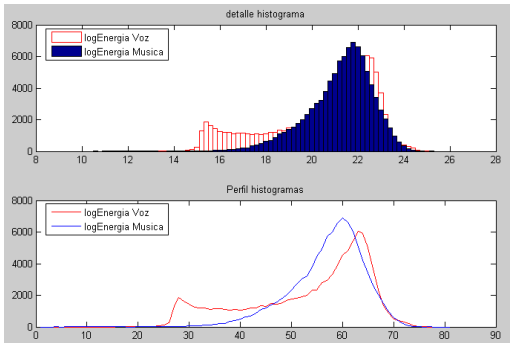

Figura 57: Histogramas logEnergia por trama

Para este caso, la distancia **KLs** es de **0.4016**.

Los resultados obtenidos son los siguientes:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	44.14	57.00	52.62	54.87	59.16	**53.56**
Cuadrático	45.62	58.25	57.20	59.34	59.67	**56.02**

Tabla 2: % aciertos de la clasificación de la LogEnergía trama a trama

Como se puede observar los resultados obtenidos son bastante mediocres, ya que unos porcentajes de acierto levemente por encima del 50% no aportan mucha información, es decir, valores entorno a este porcentaje son similares a una clasificación realizada de forma aleatoria (existen dos opciones y sin ninguna base se elije una de ellas).

5.2.2 Resultados con ZCR

La distancia KLs correspondiente a los histogramas que vemos en la figura 2 es de **0.9572**, por lo que a priori el resultado de la clasificación ha de ser un poco superior al caso analizado anteriormente.

Los resultados obtenidos para este caso son:

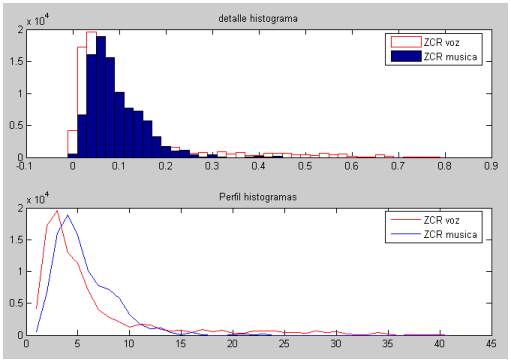

Figura 58: Histogramas ZCR por trama

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	57.51	47.38	41.44	41.57	53.79	**48.34**
Cuadrático	54.27	53.43	51.95	52.54	56.89	**53.81**

Tabla 3: % aciertos de la clasificación de ZCR trama a trama

En base solo a estos dos primeros parámetros analizados la distancia de Kullback-Leibler simétrica no guarda una relación directa con la clasificación, ya que en este caso, para una distancia KLs mayor se obtienen unos porcentajes de clasificación peores.

5.2.3 Resultados con el Logaritmo de F0

Como se aprecia en la figura de la derecha (*figura 3*) la separación entre estas dos distribuciones nos indica que la clasificación no va a poder ser todo lo buena que nos gustaría, ya que ambos histogramas se distribuyen de una manera muy similar.

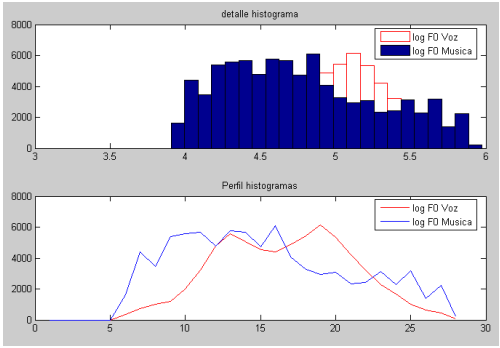

Figura 59: Histogramas logF0 por trama

En este caso la distancia KLs es **0.1092**. La distancia más baja de los tres casos analizados por tramas.

Resultados de la clasificación:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	62.26	55.00	56.49	60.95	59.47	**58.83**
Cuadrático	62.26	55.00	56.49	60.95	59.47	**58.83**

Tabla 4: % aciertos de la clasificación de LogF0 trama a trama

En este caso, se obtienen unos resultados, que aunque siguen siendo pobres, son los mejores de los tres casos analizados. También cabe destacar, que apenas encontramos diferencias entre la clasificación lineal y la cuadrática, es más, solo aparecen diferencias si analizamos hasta el cuarto decimal.

A la vista de estos resultados, se procede al análisis de la relación entre la distancia de Kullback-Leibler y el porcentaje de aciertos obtenidos:

	Distancia KLs	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
LogEnergia	0.40	53.56	56.02
ZCR	0.96	48.34	53.81
LogF0	0.11	58.83	58.83

Tabla 5: Cuadro comparativo relación distancia KLs - % Aciertos

Como puede observarse en el cuadro anterior, distancias KLs bajas producen efectivamente tasas de acierto también reducidas. Sin embargo, la distancia KLs no guarda una relación lineal con el porcentaje de aciertos en la clasificación.

Si bien es cierto que por lógica deberían guardar una relación: mayor distancia KLs – mayor porcentaje de aciertos, como se aprecia en la práctica esto no sucede. La posible razón de este comportamiento es que para el cálculo de la distancia KLs se supone que las distribuciones son Gaussianas y esto no es cierto a la vista de los histogramas correspondientes.

5.2.4 Resultados con la combinación de parámetros

Una vez analizados los parámetros de forma individual se procede a un análisis mediante la combinación de los mismos con el objetivo claro de obtener unos valores de aciertos en la clasificación considerablemente mejores, ya que, en principio, al utilizar más de un parámetro a la hora de entrenar el clasificador este tiene más criterios para elegir y clasificar correctamente.

Así pues, las combinaciones que se han llevado a cabo han sido:

- LogE+ZCR+LogF0
- LogE+LogF0

Utilizando la lógica, la primera combinación que se realiza tiene que ser la que mejor resultados ofrezca, ya que es la que más criterios de clasificación tiene, aunque como veremos a lo largo de este capitulo no siempre ocurre así. La segunda combinación la realizamos para comprobar si el parámetro ZCR ofrece algún valor extra a la clasificación o no aporta nada, ya que de forma individual sus valores apenas llegaban al 50% de aciertos en la clasificación.

Para el primero de los casos los resultados obtenidos en la clasificación son lo siguientes:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	49.96	56.55	57.79	61.62	64.51	**58.08**
Cuadrático	59.38	61.34	61.41	65.61	67.34	**63.02**

Tabla 6: % aciertos de la clasificación de LogE+ZCR+LogF0 trama a trama

Mientras que para el segundo caso de combinación de parámetros (LogE+LogF0) los resultados son:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	47.61	54.89	55.82	60.92	60.78	**56.00**
Cuadrático	54.49	57.70	58.13	65.80	60.14	**59.25**

Tabla 7: % aciertos de la clasificación de LogE+LogF0 trama a trama

Como se puede apreciar por la simple observación de ambas tablas, incluir el parámetro ZCR en la clasificación añade un 2% de aciertos en el caso de realizar una clasificación lineal, y un 4% para la clasificación cuadrática. En ninguno de los dos casos los resultados que se logran son óptimos.

A continuación se muestra una tabla comparativa de todas las clasificaciones realizadas por tramas:

Parámetros	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
LogEnergia	53.56	56.02
ZCR	48.34	53.81
LogF0	**58.83**	58.83
LogE+ZCR+LogF0	58.08	**63.02**
LogE+LogF0	56.00	59.25

Tabla 8: Cuadro comparativo % Aciertos - parámetros

Los resultados muestran que en cualquiera de los casos el porcentaje de aciertos no es lo suficientemente elevado para afirmar que con alguno de estos parámetros, o alguna combinación de los mismos, se consigue una buena clasificación.

Cabe destacar los mejores resultados que ofrece el clasificador cuadrático frente a la clasificación lineal, en torno a un 3% mejor. Esto es debido a la mayor complejidad de cómputo del clasificador cuadrático, que crea una frontera de clasificación capaz de discriminar mejor los valores que analiza. .

5.3 Clasificación con parámetros segmentales

En este apartado se van a mostrar los resultados obtenidos en la clasificación de los parámetros obtenidos en segmentos de un segundo.

En primera instancia se van a representar los resultados obtenidos al clasificar los ficheros en base a un único parámetro. Una vez vistos y analizados estos resultados se mostraran el porcentaje de aciertos al clasificar con la combinación de varios parámetros de energía, de Cruces por Cero, de Frecuencia Fundamental y combinación de parámetros de cada una de las características, con la idea final de lograr una clasificación lo más certera posible.

5.3.1 Clasificación con parámetros individuales

A continuación se muestran los histogramas obtenidos para cada uno de los parámetros y el porcentaje de aciertos de la clasificación.

Media LogEnergia

% Aciertos clasificador lineal	**60.66**
% Aciertos clasificador cuadrático	**60.03**

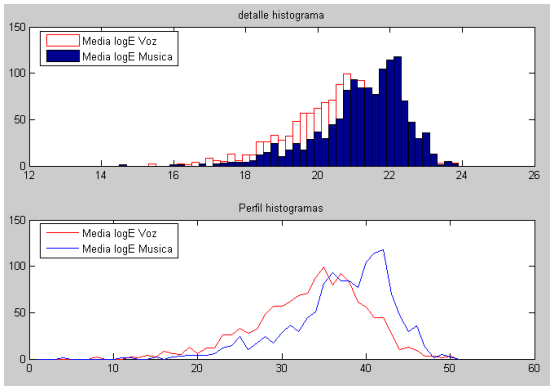

Tabla 9: Resultados e histograma Media LogEnergia

Varianza LogEnergia

% Aciertos clasificador lineal	85.12
% Aciertos clasificador cuadrático	86.28

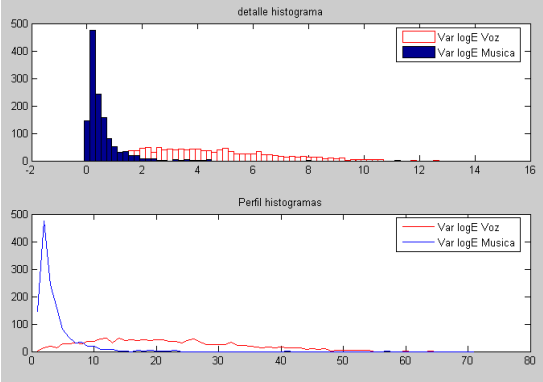

Tabla 10: Resultados e histograma Varianza LogEnergia

Porcentaje de segmentos con energía baja

% Aciertos clasificador lineal	78.41
% Aciertos clasificador cuadrático	78.53

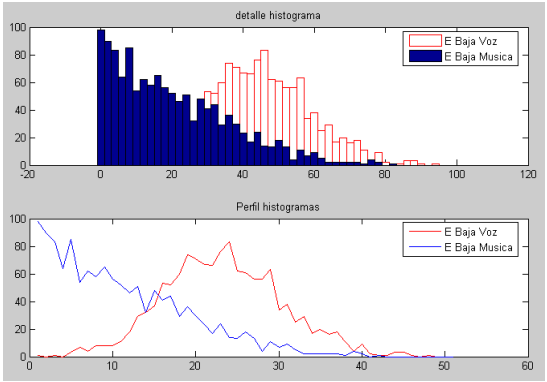

Tabla 11: Resultados e histograma Porcentaje de Segmentos con Energía Baja

Media ZCR

% Aciertos clasificador lineal	47.37
% Aciertos clasificador cuadrático	53.34

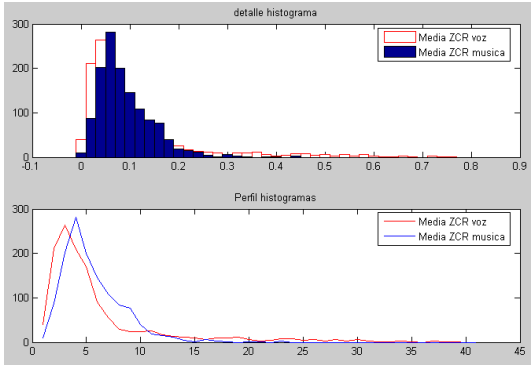

Tabla 12: Resultados e histograma Media ZCR

Varianza ZCR

% Aciertos clasificador lineal	53.06
% Aciertos clasificador cuadrático	53.28

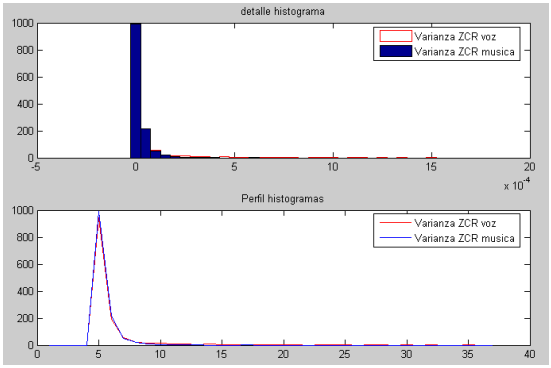

Tabla 13: Resultados e histograma Varianza ZCR

Media logF0

% Aciertos clasificador lineal	56.19
% Aciertos clasificador cuadrático	53.16

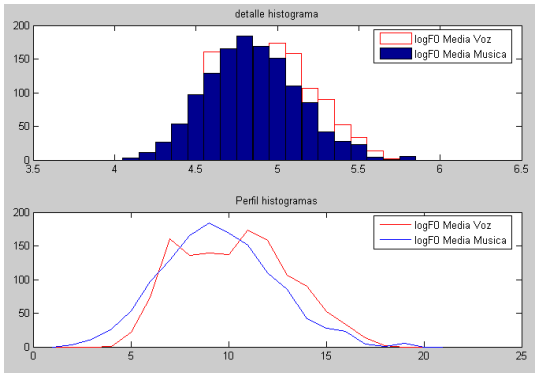

Tabla 14: Resultados e histograma Media LogF0

Varianza logF0

% Aciertos clasificador lineal	69.94
% Aciertos clasificador cuadrático	71.25

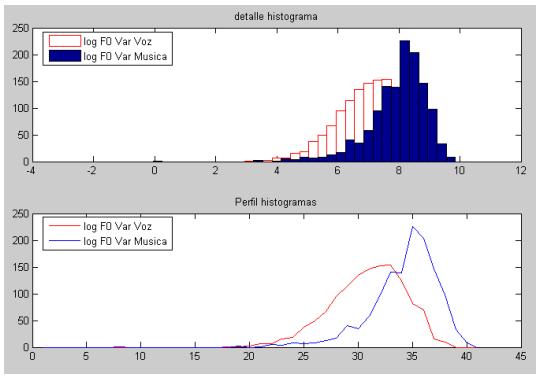

Tabla 15: Resultados e histograma Varianza LogF0

Rango de F0

% Aciertos clasificador lineal	66.00
% Aciertos clasificador cuadrático	65.56

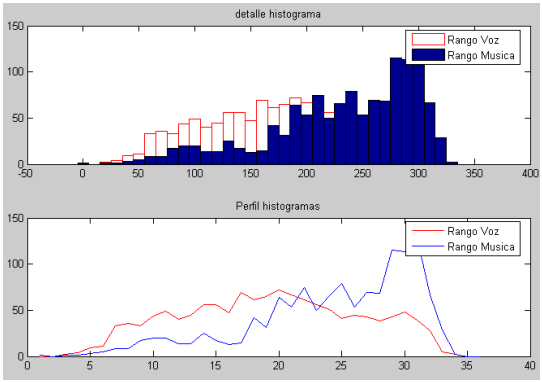

Tabla 16: Resultados e histograma Rango F0

Porcentaje de tramas sonoras

% Aciertos clasificador lineal	73.59
% Aciertos clasificador cuadrático	76.53

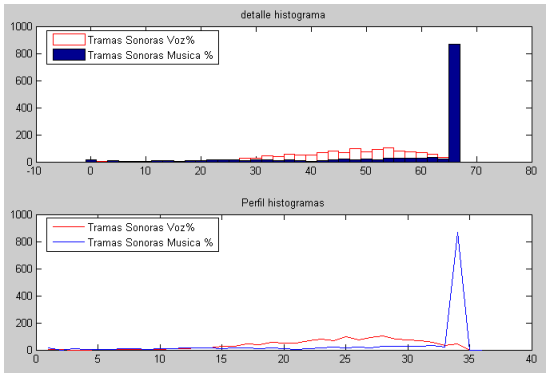

Tabla 17: Resultados e histograma Porcentaje de Tramas Sonoras

Cuadro resumen de los resultados obtenidos en parámetros individuales.

	Distancia KLs	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
Media LogEnergia	0.17	60.66	60.03
Varianza LogEnergia	9.07	85.12	**86.28**
% Energía Baja	1.30	78.41	78.53
Media ZCR	0.77	47.37	**53.34**
Varianza ZCR	2.50	53.06	53.28
Media LogF0	0.05	56.19	53.16
Varianza LogF0	0.52	69.94	71.25
Rango F0	0.28	66.00	65.56
% tramas sonoras	0.42	73.59	**76.53**

Tabla 18: Cuadro comparativo % Aciertos/Distancia KLs – parámetros

En negrita se destacan las mejores clasificaciones para cada uno de los grupos de características (energía, cruces por cero y frecuencia fundamental).

La relación que guarda la distancia KLs con el porcentaje de aciertos no es una relación lineal, como se ha comentado en el apartado anterior esto puede ser debido a que la distancia KLs parte de la premisa de que las distribuciones son gaussianas, algo que en este caso no se cumple. Si bien es cierto que las distribuciones que mayor distancia KLs presentan (‘*Varianza de la logEnergia*’) ofrecen el mejor resultado de clasificación, mientras que las que presentan una distancia KLs menor (‘*Media LogF0*’) devuelven los resultados más bajos de clasificación.

Por tanto, aunque la relación no sea siempre a mayor distancia KLs mejor clasificación y viceversa, sí que es cierto que la distancia KLs ofrece una primera idea de los resultados que se van a poder obtener.

5.3.2 Clasificación combinada con parámetros de la misma característica

En este punto se van a mostrar los resultados obtenidos en la clasificación llevada a cabo combinando diferentes parámetros. Primero dentro de una misma característica y posteriormente la combinación se realizará con parámetros de diferentes características.

Energía

Dentro de esta característica se ha optado por realizar una combinación con todos los parámetros de la misma (Media LogE + Var LogE + %Ebaja) y posteriormente se ha llevado a cabo la combinación de la 'Varianza' y del '%Ebaja', ya que son los dos parámetros que tenían un resultado independiente mejor.

El resultado obtenido para la clasificación cuando se pasan como variables los tres parámetros es el siguiente:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	87.50	88.44	90.00	93.59	88.28	**89.56**
Cuadrático	88.59	87.03	89.69	93.28	88.28	**89.37**

Tabla 19: % aciertos de la clasificación de MediaLogE+VarLogE+%Ebaja

Podemos comprobar que la mejora con respecto a utilizar solamente el mejor parámetro, la varianza, se consigue aproximadamente un 30% de reducción del error relativo en el caso del clasificador lineal y un 23% para el cuadrático. Lo que significa que el resto de parámetros aportan información extra a la clasificación.

La siguiente prueba es comprobar si realizando la clasificación sin la media, es decir, solamente con la varianza y el porcentaje de energía baja, los resultados varían significativamente. Se realiza esta comprobación ya que el

porcentaje de aciertos de la media de forma individual ronda el 60%, y es posible que no aporte nada a la clasificación combinada que acabamos de realizar.

Así pues, los resultados obtenidos cuando se pasan como variables la 'VarLogE' y el '%Ebaja' son:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	86.41	89.37	89.22	92.03	86.56	**88.72**
Cuadrático	87.81	89.69	89.84	91.87	87.19	**89.28**

Tabla 20: % aciertos de la clasificación de VarLogE+%Ebaja

Los resultados empeoran ligeramente, en la clasificación lineal apenas un 1% (0.8437%) y en la clasificación cuadrática la disminución de aciertos es aún menor (0.0937%). Por tanto, y en base a estos resultados, se puede considerar que la carga computacional que supone trabajar con un parámetro más no corresponde a una mejora suficientemente relevante.

A modo ilustrativo en la figura siguiente se observan las muestras de voz y de audio con respecto a la Varianza y al porcentaje de energía baja.

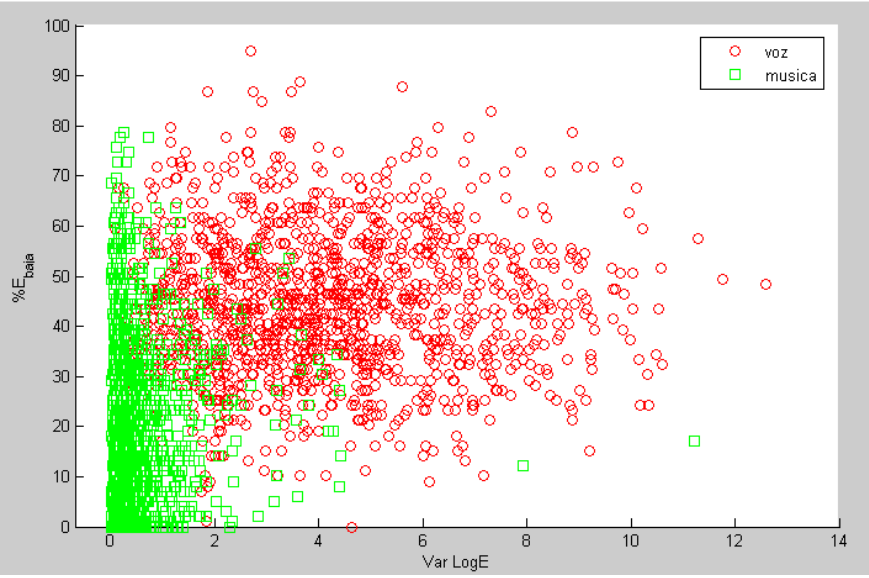

Figura 60: Dispersión de las muestras en función de la VarLogE y %Ebaja (verde = música // rojo = voz)

Cruces por cero (ZCR)

Para esta característica se realiza la única combinación posible, ya que solamente se trabaja con dos parámetros. La clasificación se realiza con la combinación de la media y la varianza de la ZCR.

El plano de dispersión que se obtiene de los archivos de voz y música en función de estos dos parámetros es el siguiente:

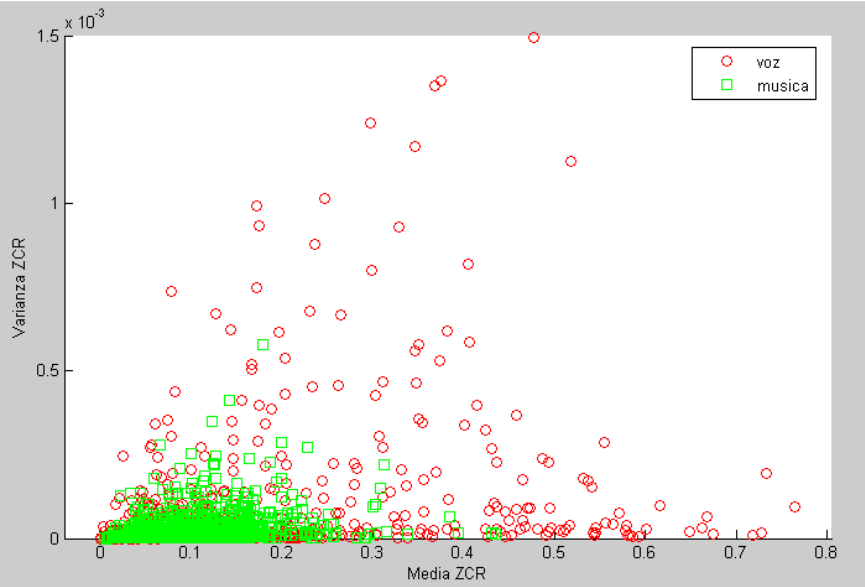

Figura 61: Dispersión de las muestras en función de la MediaZCR y VarZCR (verde = música // rojo = voz)

Como se puede observar en el gráfico superior la voz tiene una dispersión mayor, pero en ambos casos la mayor concentración de muestras está en valores bajos de media y varianza tanto para la voz como para la música, por lo que la clasificación no se espera que sea realmente buena.

Resultados de la clasificación combinada de MediaZCR y VarianzaZCR:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	58.28	49.84	45.78	48.90	58.28	**52.21**
Cuadrático	57.03	54.06	53.28	54.69	57.66	**55.34**

Tabla 21: % aciertos de la clasificación de MediaZCR+VarianzaZCR

Como era de esperar los resultados apenas superan el 50% de aciertos, por lo que este tipo de clasificación no es relevante, es como una decisión al azar.

De hecho, si se comparan estos valores con los que se obtienen al realizar la clasificación solamente con la Varianza de ZCR se puede comprobar que el porcentaje de aciertos del clasificador lineal empeora, y los aciertos del clasificador cuadrático mejoran solamente un 2%.

Así pues, este tipo de clasificación se puede desestimar para conseguir discernir entre ficheros de música y voz.

Frecuencia Fundamental

Dentro de la característica 'Frecuencia Fundamental' se trabaja con cuatro parámetros, con los cuales se van a hacer diferentes combinaciones para ver cual es el que mejor resultado ofrece con la mínima carga computacional.

Combinando los cuatro parámetros (Media del logaritmo de F0, Varianza del logaritmo de F0, Rango y porcentaje de tramas sonoras) se obtienen los siguientes resultados:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	82.81	75.94	82.81	86.72	80.47	**81.75**
Cuadrático	84.06	77.19	84.22	92.19	85.94	**84.72**

Tabla 22: % aciertos de la clasificación de MediaLogF0+VarLogF0+RangoF0+%TramasSonoras

A continuación se muestran los resultados que se obtienen pasando como parámetros al clasificador los mismos que en el caso anterior excepto la media, ya que es el que peor resultado de clasificación devuelve de forma individual:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	74.37	70.00	73.12	77.66	75.78	**74.19**
Cuadrático	80.47	70.94	74.69	83.75	82.66	**78.50**

Tabla 23: % aciertos de la clasificación de VarLogF0+RangoF0+%TramasSonoras

Se puede apreciar una clara disminución de la tasa de aciertos al quitar este parámetro, por lo que se puede concluir que realizando clasificación combinada la 'Media del logaritmo de F0' sí que aporta información relevante.

Si por el contrario realizamos la misma clasificación combinada que en el primer caso para sin el rango, ya que es el segundo parámetro que peor resultados devuelve de forma individual, los resultados son:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	82.97	76.87	83.12	85.78	79.84	**81.72**
Cuadrático	84.53	79.06	83.59	92.81	84.37	**84.87**

Tabla 24: % aciertos de la clasificación de MediaLogF0+VarLogF0+%TramasSonoras

Sorprendentemente los resultados mejoran con respecto al primer caso que se mostraba. El clasificador lineal apenas empeora un 0.0312%, mientras que el clasificador cuadrático obtiene una mejora media del 0.1562%. Ambos porcentajes no son relevantes, pero reducir una característica de trabajo proporciona un ahorro computacional claro.

La última combinación que se realiza con los parámetros de la Frecuencia Fundamental es utilizar para la clasificación la 'Varianza del logaritmo de F0' y el 'Porcentaje de tramas sonoras', que son los dos parámetros que de forma individual ofrecen un mejor resultado.

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	77.81	70.00	73.59	79.69	75.78	**75.37**
Cuadrático	82.19	75.62	74.53	86.25	82.19	**80.16**

Tabla 25: % aciertos de la clasificación de VarLogF0+%TramasSonoras

El diagrama de dispersión asociado a esta clasificación es el siguiente:

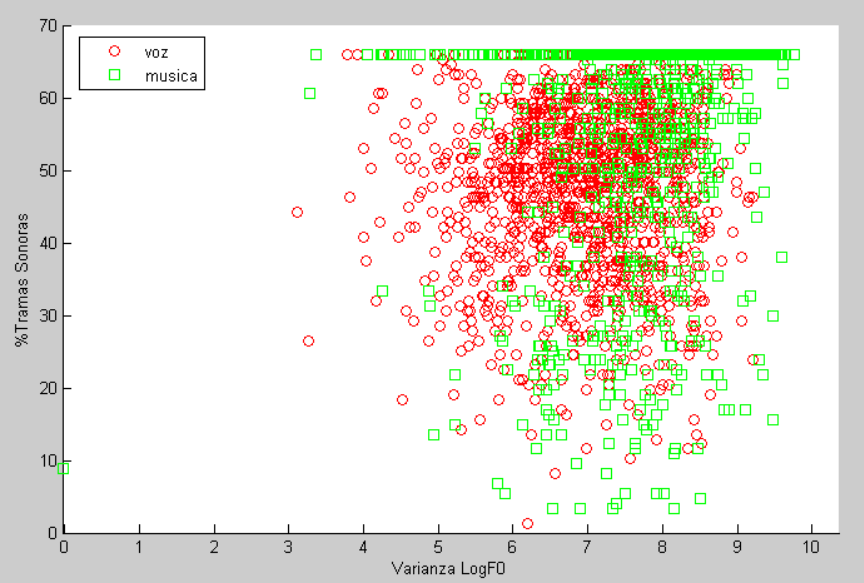

Figura 62: Dispersión de las muestras en función de VarLogF0 y %TramasSonoras (verde = música // rojo = voz)

A continuación se muestra un cuadro resumen de los resultados obtenidos en la clasificación con la combinación de parámetros de la Frecuencia Fundamental.

	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
MediaLogF0+VarLogF0+ +Rango+%TramasSonoras	**81.75**	84.72
VarLogF0+Rango+ +%TramasSonoras	74.19	78.50
MediaLogF0+VarLogF0+ +%TramasSonoras	81.72	**84.87**
VarLogF0+%TramasSonoras	75.37	80.16

Tabla 26: Cuadro comparativo % Aciertos – parámetros F0

Si se analizan los resultados reflejados en el cuadro superior vemos varias cosas que hay que tener en cuenta:

Como ya se ha dicho anteriormente la media es relevante, ya que si se suprime la clasificación empeora significativamente.

Por otra parte, si lo que se quita en el rango la tasa de aciertos apenas varía (la clasificación lineal empeora y la cuadrática mejora, en ambos casos muy ligeramente).

Por último, si se realiza la clasificación con la varianza, el rango y el porcentaje de tramas sonoras se obtiene un resultado peor que si solamente se utiliza la varianza y el porcentaje de tramas sonoras, de lo que se puede concluir que el rango no aporta información a la hora de clasificar, es más, produce una degradación en los resultados.

A continuación se muestra una tabla resumen con todos los resultados de las diferentes clasificaciones que se han presentado en este apartado:

	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
MediaLogE+VarLogE+%Ebaja	**89.56**	**89.56**
VarLogE+%Ebaja	88.72	88.72
MediaZCR+VarZCR	52.21	52.21
MediaLogF0+VarLogF0+ +Rango+%TramasSonoras	81.75	84.72
VarLogF0+Rango+ +%TramasSonoras	74.19	78.50
MediaLogF0+VarLogF0+ +%TramasSonoras	81.72	84.87
VarLogF0+%TramasSonoras	75.37	80.16

Tabla 27: Cuadro comparativo % Aciertos – parámetros

Como se puede observar en el cuadro superior la combinación de parámetros que proporciona una tasa de aciertos mayor es la que se consigue con los parámetros de la energía. Por el contrario los parámetros ZCR ofrecen un porcentaje de aciertos bastante mediocre que apenas aportará nada al conjunto global de la clasificación.

5.3.3 Clasificación combinada de parámetros de diferentes características

Para lograr unos resultados óptimos en la clasificación es lógico pensar que combinando los mejores parámetros de cada una de las características se conseguirá. Del mismo modo que es lógico que si se utilizan todos los parámetros que se han desarrollado en los apartados anteriores se obtenga la mejor clasificación, ya que cuanta más información se pase al clasificador mejor resultado devolverá.

Por tanto el experimento de clasificar con todos los parámetros que se han desarrollado ofrece el siguiente resultado:

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	88.75	91.25	93.12	97.03	91.72	**92.37**
Cuadrático	89.53	85.47	92.66	96.09	91.25	**91.00**

Tabla 28: % aciertos de la clasificación con todos los parámetros analizados

Los resultados que se obtienen son realmente buenos, superior al 90% de aciertos tanto para una clasificación con frontera lineal como con frontera cuadrática.

Pero teniendo los resultados de los apartados anteriores presentes se puede llegar a la conclusión de que ciertos parámetros es posible que no aporten nada de información o muy poca, por lo que serán suprimibles, ahorrando de este modo carga computacional.

Por todo ello se realizaran clasificaciones con todos los parámetros excepto el Rango de F0 por una parte; excepto las dos características de ZCR por otra. Se realiza la clasificación sin el parámetro rangoF0 para comprobar si tiene un efecto negativo en el cómputo global de la clasificación con todos los parámetros, ya que en la clasificación con parámetros F0 empeoraba el clasificador. Se lleva a cabo la clasificación sin los parámetros de ZCR ya que

estos de forma individual apenas alcanzaban el 50% de acierto. Además se realiza una clasificación suprimiendo los parámetros de ZCR y el rango de F0 para comprobar la tasa de aciertos que devuelve.

Por último se lleva a cabo un planteamiento inverso, en lugar de partir de todas las características y quitar las peores para ver como se va manteniendo el resultado, se seleccionan las características que de forma individual han ofrecido más de un 70% de aciertos para ver si combinándolas se consigue una clasificación con resultados óptimos. Las cuatro características seleccionadas para esta clasificación son la 'Varianza del logaritmo de F0', el 'porcentaje de tramas sonoras' (parámetro asociado a la F0), el 'porcentaje de energía baja' y la 'varianza del logaritmo de la energía'.

El resumen de los resultados que se obtienen en este apartado de "Clasificación combinada de parámetros de diferentes características" se muestra a continuación:

	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
Todos los parámetros	**92.37**	91.00
Todos los parámetros excepto rangoF0	91.87	90.91
Todos los parámetros excepto ZCR	92.31	**93.12**
Todos los parámetros excepto ZCR y RangoF0	91.81	92.53
VarLogF0+%TramasSonoras+ +%Ebaja+VarLogE	89.31	91.84

Tabla 29: Cuadro comparativo % Aciertos – parámetros inter-características

En el cuadro anterior se muestra en negrita los mejores resultados de clasificación para el clasificador lineal y el cuadrático.

Los resultados obtenidos son bastante representativos, pero cabe destacar algún detalle:

- En la clasificación con solamente cuatro parámetros el clasificador lineal empeora alrededor de un 3%, que corresponde con un incremento del error relativo de aproximadamente el 28%. En este caso sería conveniente utilizar todas las características acústicas. Sin embargo, con el clasificador cuadrático, la reducción de parámetros conlleva una reducción del error relativo de en torno al 10%. En ambos casos los resultados son suficientemente válidos, por lo que es posible que estos cuatro parámetros sean los que mejor resultado ofrecen.
- Por el contrario hay que destacar de forma negativa los parámetros de ZCR, ya que realizando la clasificación sin ellos se obtiene un resultado mejor con el clasificador cuadrático y un resultado similar con el lineal, por lo que estos parámetros no ofrecen ningún criterio de uso para la clasificación. Si bien es cierto que no se esperaba una tasa de aciertos tan baja en base a los resultados obtenidos en el '*Capítulo 3: Caracterización paramétrica de música y voz*'.

5.3.4 Comparación trama a trama/segmental

En este apartado se va a comparar los resultados globales de realizar una clasificación con parámetros obtenidos por tramas y la clasificación con parámetros segmentales.

A continuación se muestra la tabla comparativa:

	% Aciertos Clasificador lineal	% Aciertos Clasificador cuadrático
LogF0 por Trama	**58.83**	58.83
LogE+ZCR+LogF0 por Trama	58.08	**63.02**
LogE+LogF0 por Trama	56.00	59.25
Todos los parámetros segmentales	**92.37**	91.00
Todos los parámetros segmentales excepto rangoF0	91.87	90.91
Todos los parámetros segmentales excepto ZCR	92.31	**93.12**
Todos los parámetros segmentales excepto ZCR y RangoF0	91.81	92.53
VarLogF0+%TramasSonoras+ +%Ebaja+VarLogE por Segmento	89.31	91.84

Tabla 30: Cuadro comparativo % Aciertos – parámetros inter-características

En negrita se muestran los mejores resultados de la clasificación por tramas y por segmentos. A su vez, los valores que están en negrita y subrayados indican la tasa de aciertos más alta en cualquiera de los casos.

Como se aprecia en el cuadro, la tasa de aciertos que se obtiene cuando el clasificador trabaja con parámetros obtenidos por tramas es muy inferior a la que se obtiene con parámetros segmentales.

El cálculo de los parámetros segmentales lleva una carga computacional ligeramente mayor, pero que compensa con creces el resultado que ofrece. Además, no hay que olvidar, que la idea de trabajar con parámetros segmentales es por que de este modo nos asemejamos más la integración que hace el oído de los sonidos, por lo que en conjunto analizar estos parámetros se acerca más a la realidad.

Capítulo 6. Conclusiones y líneas futuras

6.1 Introducción

En este capítulo se considera todo lo que se ha observado durante el desarrollo del proyecto, y sobre todo se extraen conclusiones a partir de los resultados experimentales que se han ido mostrando a lo largo del capítulo anterior. También se mostrarán algunas líneas de investigación que podrían servir como trabajo en un futuro.

6.2 Conclusiones

Como se ha comentado a lo largo de los capítulos anteriores, para la clasificación de audio en dos clases (música instrumental y voz) trabajamos con tres grupos de características (la energía, la tasa de cruces por cero y la frecuencia fundamental).

Revisando los resultados obtenidos y que se reflejan en el capítulo anterior podemos sacar varias conclusiones:

- La tasa de cruces por cero (ZCR) no aporta un valor añadido a la clasificación, es más, es mejor no considerar esta característica a la hora de clasificar. Se obtienen mejores resultados de clasificación si no se utiliza esta tasa de cruces por cero (ZCR), por lo que no trabajar con ella ahorra carga computacional y da mejores resultados.
- El clasificador cuadrático, de forma general, ofrece mejores resultados de clasificación que el clasificador lineal. Aunque las diferencias son pequeñas, hay que tener en cuenta que computacionalmente es más complejo que el clasificador lineal.

- Podemos afirmar que los mejores parámetros para realizar la clasificación, en relación carga computacional – resultado obtenidos, son: la varianza del logaritmo de F0, el porcentaje de tramas sonoras, el porcentaje de tramas con energía baja y la varianza del logaritmo de la energía. Los dos primeros parámetros corresponden a la frecuencia fundamental, mientras que los otros dos son parámetros de la energía.
- Los resultados que se obtienen en la clasificación con parámetros extraídos a nivel de segmento son notablemente mejores que la tasa de aciertos que se logra con los parámetros extraídos por trama.

6.3 Líneas futuras de trabajo

Es difícil definir de una forma clara posibles líneas de trabajo, ya que los campos que se abren a nuestro alrededor son tan grandes como la imaginación humana. Así pues, a continuación solo se va a hacer mención de unas posibles líneas, las más obvias y cercanas a lo que en este proyecto se ha analizado.

- Habría que trabajar con un conjunto de ficheros mayor y grabaciones de muestras en diferentes escenarios (por ejemplo con ruido) para así obtener unos resultados aplicables a diferentes aspectos tecnológicos que puedan surgir.
- Considerar nuevos conjuntos de parámetros acústicos (como MFCC, energía en banda, etc.) y su combinación con los utilizados en este proyecto.
- Investigar tareas más complicadas como puede ser el reconocimiento de los estilos musicales, para lo que se necesitaría una biblioteca de clases más amplia, y una parametrización más compleja.
- Extender el trabajo a la detección de eventos acústicos (gritos, accidentes, goles, calma…).
- Reconocedores de cantantes en función de los parámetros analizados a lo largo de este proyecto.

Capítulo 7. Presupuesto

7.1 Introducción

Una parte fundamental a la hora de desarrollar un proyecto de ingeniería es hacer una estimación de los costes económicos que supondría su desarrollo. Estos van a marcar la viabilidad económica del mismo y muy seguramente determinarán la realización o no del proyecto.

En el presente capítulo se presenta un cálculo aproximado del presupuesto, en el que se han considerado dos tipos de coste fundamentales: los derivados de los honorarios de los desarrolladores y los derivados de los costes materiales.

7.2 Honorarios de los desarrolladores

Los costes de honorarios comprenden todos aquellos aspectos que no están relacionados con la compra de equipos, licencias, etc. Dentro de estos costes se tienen en cuenta dos conceptos diferentes:

Costes asociados al desarrollo del proyecto

Tomando como referencia el informe del Colegio Oficial de Ingenieros Técnicos de Telecomunicaciones (COITT) “Baremos de Honorarios Orientativos para Trabajos Profesionales”, los costes para “Trabajos por tiempo empleado” son:

- 60 €/hora, para las horas normales (No se van a considerar en ningún caso horas extraordinarias)

La dedicación empleada para la realización de este proyecto, se puede estimar en 15 horas a la semana durante 8 meses, lo que supone la cantidad de

480 horas. Los honorarios en concepto de desarrollo del proyecto ascienden a un valor VEINTIOCHO MIL OCHOCIENTOS EUROS (28.800 €) sin IVA.

Costes asociados a la dirección del proyecto

Para calcular los costes asociados a la dirección del proyecto se aplica el coeficiente estipulado por el COITT. El coste se obtiene como el 7 % del coste de desarrollo del proyecto, lo que asciende a un valor de DOS MIL DIECISEIS EUROS (2.016 €) sin IVA.

7.3 Costes materiales

Los costes materiales comprenden todos aquellos aspectos que están relacionados con la compra de equipos, licencias, etc. En la siguiente tabla se detallan los costes de los equipos empleados, los costes de las licencias empleadas, así como los costes de alquiler de la oficina empleada.

Concepto	Coste
Ordenador Personal de sobremesa	500 €
Software Matlab 7.0 con licencia para uso personal	320 €
Material de oficina y no tangible	275€
Local (8 meses/145€mes)	1160 €
TOTAL	**2255€**

Tabla 31: Relación de conceptos y sus costes

Los costes de oficinas agrupan diferentes conceptos como fotocopias, compra de libros, documentación, etc.

7.4 Presupuesto total

Sumando los costes por honorarios por desarrolladores y los costes materiales se obtiene que el presupuesto total del proyecto asciende a TREINTA Y TRES MIL SETENTA Y UN EUROS (33.071 €) sin IVA.

Aplicando un coeficiente de IVA del 16 %, se obtiene que el montante total del presupuesto asciende a un valor de TREINTA Y OCHO MIL TRESCIENTOS SESENTA Y DOS EUROS CON TREINTA Y SEIS CÉNTIMOS (38.362,36 €).

Capítulo 8. Referencias

[2.1] "Speech/Music discrimination for multimedia applications"
Khaled El-Maleh, Mark Klein, Grace Petrucci, Peter Kabal
IEEE International Conference on Acoustics, Speech and
Signal Processing (ICASSP'00), vol. 4, pp. 2445-2448, 2000.

[2.2] "An Overview of Speech/Music Discrimination Techniques in the Context of Audio Recordings"
Aggelos Pikrakis, Theodoros Giannakopoulos, and Sergios Theodoridis
Multimedia Services in Intelligent Environments (Capítulo 4), pp. 81-102, Ed. Springer, 2008.

[2.3] "Speech/music segmentation using entropy and dynamism features in a HMM classification framework"
Jitendra Ajmera, Iain McCowan, Hervé Bourlard
Speech Communication, vol. 40, pp. 351-363, 2003.

[3.1] "Detector de actividad de voz basado en la distancia de Kullback-Leibler con la aplicación a reconocimiento robusto de voz"
Javier Ramírez, José C. Segura, Carmen Benítez, Ángel de la Torre, Antonio J. Rubio
Departamento de Electrónica y Tecnología de Computadores
Universidad de Granada

[3.2] Wikipedia. La enciclopedia libre
http://es.wikipedia.org/wiki/Energ%C3%ADa

[3.3] "Análisis localizado en tiempo y frecuencia"
Tratamiento de Señales Audiovisuales
Tratamiento digital de la señal de voz
http://agamenon.tsc.uah.es/Asignaturas/it/tdv/apuntes/TSA_AnaLoc.pdf

[3.4] "Análisis localizado de voz"
Universidad Autónoma de Madrid
http://arantxa.ii.uam.es/~jortega/Tema2_TAPS_def.pdf

[3.5] "Clasificación automática de fuentes de ruido de tráfico"
Manuel A. Sobreira Seoane, Alfonso Rodríguez Molares
Sonitum®, E.T.S.I. de Telecomunicación, Universidad de Vigo.

[3.6] "Clasificación automática de voz/música utilizando discriminantes lineales de Fisher"
Enrique Alexandre Cortizo, Manuel Rosa Zurera
Departamento de Teoría de la Señal y Comunicaciones
Universidad de Alcalá

[3.7] "Análisis de la señal de voz"
Asignatura: Audio digital II. Universidad de Extremadura
http://tsc.unex.es/~ycampos

[3.8] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[3.9] "EE214A Project: Pitch, Formant Tracking Systems"
Huiyu Luo, Rick Lan, Joshua Cantrell and Nima Kazemi
Department of Electrical Engineering, UCLA

[4.1] Wikipedia. La Enciclopedia libre
http://es.wikipedia.org/wiki/Clasificadores_(matem%C3%A1tico)

[4.2] "Analizador numérico" Fernando Berzal.
Intelligent Databases and Information Systems research group
Department of Computer Science and Artificial Intelligence
E.T.S Ingeniería Informática – Universidad de Granada (Spain)

[4.3] " Reconocimiento de Formas en Data Mining" Héctor Allende
www.inf.utfsm.cl

Proyectos de fin de carrera que se han tenido presentes para la ejecución de este trabajo:

- "Implementación de un verificador de hablante independiente del texto para dispositivos móviles Symbian" Autor: Felipe Díaz Frutos. Tutor: Fernando Díaz de María

- "Técnicas de normalización de parámetros cepstrales para reconocimiento robusto de habla" Autor: Luis Elvira Álvarez. Tutora: Ascensión Gallardo Antolín

- "Estudio comparativo de parámetros espectrales para clasificación de audio" Autor: Enrique Prieto Labrador. Tutora: Ascensión Gallardo Antolín

Anexo. Resultados experimentales

1 Introducción

En este anexo se incluirán todos los resultados que de forma experimental se han obtenido a lo largo de este proyecto, lo que permite tener todos los resultados agrupados para un estudio de los mismos más sencillo y directo.

Los resultados, al igual que en el capítulo 5, se presentan en una tabla, donde los valores identificados en las casillas del grupo 1 al grupo 5 indican la tasa de aciertos de las cinco rotaciones de la validación cruzada con el clasificador lineal y del clasificador cuadrático. El valor que se encuentra a la derecha ('Media') indica los valores medios de la validación cruzada en cada uno de los casos.

2. Resultados con parámetros obtenidos por tramas (% de aciertos)

Clasificación LogEnergia

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	44.14	57.00	52.62	54.87	59.16	**53.56**
Cuadrático	45.62	58.25	57.20	59.34	59.67	**56.02**

Clasificación Cruces por cero

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	57.51	47.38	41.44	41.57	53.79	**48.34**
Cuadrático	54.27	53.43	51.95	52.54	56.89	**53.81**

Clasificación Logaritmo de la Frecuencia Fundamental

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	62.26	55.00	56.49	60.95	59.47	**58.83**
Cuadrático	62.26	55.00	56.49	60.95	59.47	**58.83**

Clasificación LogE+ZCR+LogFo

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	49.96	56.55	57.79	61.62	64.51	**58.08**
Cuadrático	59.38	61.34	61.41	65.61	67.34	**63.02**

Clasificación LogE+LogF0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	47.61	54.89	55.82	60.92	60.78	**56.00**
Cuadrático	54.49	57.70	58.13	65.80	60.14	**59.25**

3. Resultados con parámetros obtenidos por segmentos (% aciertos)

Clasificación Media LogEnergia

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	52.66	66.40	60.94	60.16	63.12	**60.66**
Cuadrático	51.41	66.09	57.34	57.03	68.28	**60.03**

Clasificación Varianza LogEnergia

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	80.16	82.81	90.62	89.69	82.34	**85.12**
Cuadrático	83.59	83.59	90.78	89.68	83.75	**86.28**

Clasificación Porcentaje Energía Baja

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	78.75	80.62	72.97	82.97	76.72	**78.41**
Cuadrático	79.06	81.09	72.19	83.44	76.87	**78.53**

Clasificación MediaZCR

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	56.56	45.31	38.75	42.03	54.22	**47.37**
Cuadrático	55.16	50.94	50.31	52.81	57.50	**53.34**

Clasificación Varianza ZCR

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	58.90	52.19	47.97	50.16	56.09	**53.06**
Cuadrático	56.41	52.34	51.25	51.72	54.68	**53.28**

Clasificación Media LogF0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	54.22	58.12	62.50	59.53	46.56	**56.19**
Cuadrático	50.78	60.62	56.72	49.84	47.81	**53.16**

Clasificación Varianza LogF0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	69.84	66.72	67.97	72.03	73.12	**69.94**
Cuadrático	74.53	69.69	65.31	72.19	74.53	**71.25**

Clasificación Rango F0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	64.69	63.12	65.94	68.90	67.34	**66.00**
Cuadrático	63.91	62.66	65.31	68.75	67.19	**65.56**

Clasificación Porcentaje de Tramas Sonoras

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	79.37	68.12	70.78	76.72	72.97	**73.59**
Cuadrático	82.03	69.53	72.66	81.87	76.56	**76.53**

Clasificación MediaLogE+VarLogE+%Ebaja

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	87.50	88.44	90.00	93.59	88.28	**89.56**
Cuadrático	88.59	87.03	89.69	93.28	88.28	**89.37**

Clasificación VarLogE+%Ebaja

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	86.41	89.37	89.22	92.03	86.56	**88.72**
Cuadrático	87.81	89.69	89.84	91.87	87.19	**89.28**

Clasificación MediaZCR+VarZCR

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	58.28	49.84	45.78	48.90	58.28	**52.21**
Cuadrático	57.03	54.06	53.28	54.69	57.66	**55.34**

Clasificación MediaLogF0+VarLogF0+RangoF0+%TramasSonoras

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	82.81	75.94	82.81	86.72	80.47	**81.75**
Cuadrático	84.06	77.19	84.22	92.19	85.94	**84.72**

Clasificación VarLogF0+RangoF0+%TramasSonoras

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	74.37	70.00	73.12	77.66	75.78	**74.19**
Cuadrático	80.47	70.94	74.69	83.75	82.66	**78.50**

Clasificación MediaLogF0+VarLogF0+%TramasSonoras

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	82.97	76.87	83.12	85.78	79.84	**81.72**
Cuadrático	84.53	79.06	83.59	92.81	84.37	**84.87**

Clasificación VarLogF0+%TramasSonoras

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	77.81	70.00	73.59	79.69	75.78	**75.37**
Cuadrático	82.19	75.62	74.53	86.25	82.19	**80.16**

Clasificación todas las características segmentales

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	88.75	91.25	93.12	97.03	91.72	**92.37**
Cuadrático	89.53	85.47	92.66	96.09	91.25	**91.00**

Clasificación Todas las Características excepto RangoF0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	89.22	89.53	92.34	97.03	91.25	**91.87**
Cuadrático	89.84	85.78	92.34	96.09	90.47	**90.91**

Clasificación todas las características excepto parámetros ZCR

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	88.44	91.25	93.44	97.18	91.25	**92.31**
Cuadrático	90.16	90.16	95.94	96.87	92.50	**93.12**

Clasificación todas las características excepto parámetros ZCR y RangoF0

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	89.22	89.37	92.81	96.87	90.78	**91.81**
Cuadrático	89.37	89.53	95.94	96.25	91.56	**92.53**

Clasificación VarLogF0+%TramasSonoras+%Ebaja+VarLogE

Clasificador	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Media
Lineal	85.47	88.91	90.31	93.28	88.59	**89.31**
Cuadrático	88.90	91.71	91.25	95.31	92.03	**91.84**